

Tea but no cake

As Japan strives to encourage academics to work more closely with industry, progress is hindered by strict government regulations, researchers' fear of criticism and industry's reluctance to fund research that may not bring profits.

For years, officials in Japan's Ministry of International Trade and Industry (MITI) and their somewhat less enthused colleagues in the Ministry of Education, Science and Culture (Monbusho) have been trying to encourage academics to work more closely with industry.

Earlier this year, regulations were eased to allow university faculty members to assume executive positions in industry (see *Nature* **403**, 589; 2000), and now a new fund has been created to encourage interaction between the universities and industry (see News, pages 925–926). But if these measures are to have any effect, some deeply embedded attitudes — reflected in ideas about patenting and in some seemingly petty regulations — will have to be overcome.

As public employees, university researchers often hesitate to join hands with industry, fearing criticism that they are seeking to profit from research done at the public's expense. At the same time, individual researchers are hesitant to put in all the hard work needed to produce a proper patent for industrial use if they cannot profit from it. Under current regulations, short of going through what many consider a prohibitively laborious evaluation process, all patents and royalties revert to the university, not to the individual.

For its part, industry is loath to fork out money for university research when patents are unlikely to be filed promptly or written with great enough breadth to promise high returns. It is also concerned that patents are going to be tied to conservative public institutions, rendering their use or sale dependent on lengthy, bureaucratic deliberations. In the current state of play, there is "no hope that we can win in the patent war", one official laments.

Then there is what might be termed 'the cake problem'. University and national institute researchers are being encouraged to work with

industry, but strict regulations laid down this March (the same month as those allowing university faculty members to assume executive positions in industry) limit their personal interaction. As one frustrated researcher puts it: "We can't even have cake together. Tea is probably OK. It's hard to tell exactly." Such regulations even prohibit university lecturers from having lunch with graduate students, lest their ability to evaluate them objectively is compromised.

The background to these regulations lies in Japan's unfortunate recent history of golf-club memberships and fancy dinners being used to bribe public officials. The blanket coverage of all public employees, including university staff, may seem arbitrary. But legislators probably fear that publicly funded researchers (and thus public resources) are vulnerable in their collaborations with industrial partners — a criticism that arises even of the university–industry partnerships that are now well established in the United States.

Although recession has cut into Japan's vaunted industrial research sector, university–industry collaborations have the potential to play an important role in sustaining Japan's industrial might. But over-zealous efforts to protect researchers from being tainted by contacts with commerce are likely to foil the government's efforts to bring industry and the universities together.

Academics' lack of familiarity with the patent system is an additional impediment to progress. Setting up some technology licensing offices at the universities could go a long way towards addressing that problem. But, if scientists are going to respond actively and enthusiastically to its initiatives, the government will have to coordinate its efforts better and devise a workable regulatory framework for interaction between universities and industry. ■

Awkward inconsistencies of a stem-cell rule

The US government's clever interpretation of the law lets stem-cell research proceed, but leaves it exposed to challenges.

The revised guidelines for funding human embryonic stem-cell research published by the US National Institutes of Health last week provide an imperfect solution to an insoluble problem. The research potential of the cells demands that NIH-funded biomedical researchers should be allowed to use them. Embryonic stem cells can divide for ever and, with work, could be steered into becoming virtually any other cell type. Harnessing stem cells could one day provide treatments for spinal cord injury, neurodegenerative diseases and diabetes. But many in the United States do not believe this outweighs the moral price of deriving research materials from discarded human embryos.

The guidelines allow federal funds for stem-cell research but not for the derivation of the cells themselves. They satisfy US researchers' immediate concerns, but they rest on fragile logic. Researchers who receive federal funds to study the cells will end up using federal grant money to pay those who derive the cells anyway. Opponents of abortion correctly point out that the rules do not clarify existing laws against embryonic research, but rather circumvent them. They say funding experimentation tacitly supports derivation, and the destruction of embryos which that entails.

Researchers on federal grants worry that privately funded colleagues will gain an edge in experimental manipulation of stem cells which they derive for themselves. NIH-funded researchers will be limited to the cell lines provided by private firms, rather than being free to create and tailor lines to meet their own needs. Scientists in Britain and elsewhere will probably gain permission to both derive and experiment on embryonic stem cells (see *Nature* **406**, 815; 2000).

Reliance on the distinction between use and derivation leaves the future of the research uncertain, subject to the influence of politics and of the courts. If vice-president Al Gore wins the US presidential election, he will be inclined to let it continue under the NIH guidelines. But if George W. Bush prevails, he could easily pass an executive order banning all federally funded stem-cell research.

A bill proposed by Senator Arlen Specter (Republican, Pennsylvania) to allow funding for use and derivation of embryonic stem cells would end the ambiguity. But the bill hasn't yet been brought to a vote in the Senate and will struggle to muster support in the House of Representatives. Progress must await the outcome of the presidential and Congressional elections in November. ■

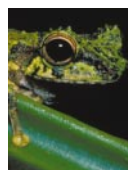




Paper chase
Cambridge sets
its sights on
Newton archive
p924



Stem cells
NIH clears the way
for research on
embryonic material
p925



Hotspots
Ecologists discuss
priorities for
bioconservation
p926



For the chop?
Ammonia causes
a stink in
North Carolina
p928

Wellcome discusses structural genomics effort with industry...

Declan Butler

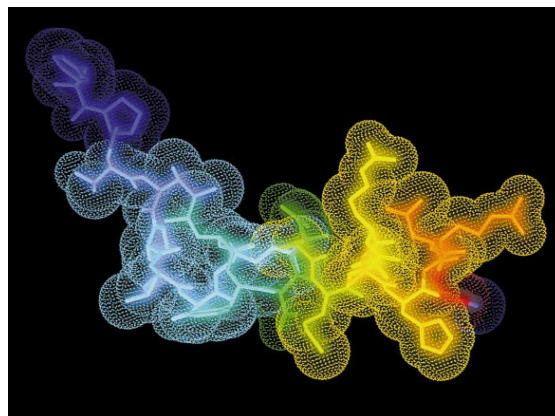
Britain's Wellcome Trust and several companies are discussing the creation of an international consortium to advance structural genomics — the determination of the structures of proteins, RNA and other biological macromolecules.

Currently, most structures are deduced slowly and expensively by researchers studying a particular macromolecule. With the draft sequence of the human genome now available, the consortium's eventual goal would be to complement this approach with an industrial-scale drive to develop the high-throughput determination of thousands of structures.

Alan Williamson, former vice-president for worldwide research strategy at Merck and now an independent consultant, is championing the project. An interim committee has been created to study its feasibility.

Williamson played an important role in brokering the so-called SNP Consortium, a non-profit foundation that aims to put 300,000 single-nucleotide polymorphisms (SNPs) — differences in single DNA base pairs between individuals — into the public domain by the end of 2001.

The companies on the interim committee



Shape of things to come? An industrial-scale attack on protein structures — such as the growth hormone somatotropin, shown here — could aid drug design.

have not been named. But some of the 13 who are members of the SNP consortium — including IBM, one of the latest recruits — are believed to be involved.

"An effort to significantly increase the number of target protein structures in the public domain would be of great interest to anyone involved in rational drug design," says Rob Cook, head of molecular recognition at one of these companies, Glaxo-Wellcome in Stevenage, England.

"I was approached by several companies interested in increasing the rate [at which]

structures flow into public-domain databases," says Williamson. "Now we want to open up the consortium to others." A decision on whether to go ahead might be taken by the end of the year, he says.

The cost of consortium membership has not been fixed, but is likely to be around US\$1.5 million per year. Structures would be determined at public research institutes.

"We certainly see the consortium as potentially a very good way to further structural genomics," says Michael Morgan, chief executive of the Wellcome Trust's Genome Campus near Cambridge, England. "We have an 'in principle' agreement with the governors of the trust that if we can get the consortium to work, [the fund] will chip in to the level needed to make it a success."

The move reflects growing academic interest in taking structural genomics to an industrial scale, although views differ on when and how this should be done.

At a meeting in Cambridge in April, organized by the US National Institute of General Medical Sciences (NIGMS) and the Wellcome Trust, scientists from public agencies in nine countries agreed to form an 'International Structural Genomics Initiative'.

This initiative aims to coordinate national activities, avoid duplication of effort, and adopt common data-release policies. Its goal is to produce 10,000 protein structures in ten years.

But the initiative does not dictate targets and strategy or fund structural centres. It

...but data release remains an open question

One of the question marks hanging over cooperation in structural genomics is how and when the structures of proteins and other macromolecules should be made public.

At a meeting in Cambridge in April, organized by the Wellcome Trust, scientists from public agencies in nine countries agreed that there should be timely release of all data.

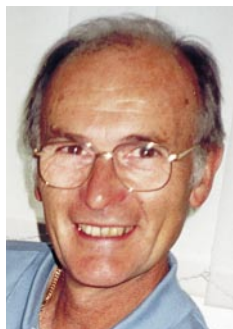
But a loophole remains in that, for the next couple of years, the decision on when a structure is complete and ready for a

database is left to researchers themselves. This is because there is no agreement or way of defining when a protein structure is accurate and complete — it is largely a matter of personal judgement.

Arriving at an automatic cut-off point based on "numerical criteria" of quality is "now a key goal", says John Norvell, head of the US Protein Structure Initiative based at the National Institute of General Medical Sciences. This would allow for clear rules on the timing of data release.

An early agreement between researchers on data access is also critical to facilitating collaboration with industry. As several delegates at a meeting organized by the Organisation for Economic Cooperation and Development in Florence in June pointed out, researchers may be reluctant to put valuable structural data into the public domain.

The Cambridge meeting agreed that data released should be accompanied by a short research paper, published simultaneously on the Internet. **D. B.**



Williamson: champion of structural consortium.

may evolve to do this, says John Norvell, head of the US Protein Structure Initiative. But before this can happen, many technical and logistical issues need to be resolved.

The Protein Structure Initiative hopes to deal with these issues; it will announce its first round of grants to a handful of pilot centres in the United States next month.

The three existing large programmes in structural genomics are run by the NIGMS, the Institute of Physical and Chemical Research (RIKEN) in Japan, and the Wellcome Trust.

Norvell agrees that a consortium could speed things up. "The participation of industry in the international effort is strongly encouraged," he says. But the NIGMS has not joined the consortium. "It's early in the game for all of us," says Norvell.

The Wellcome Trust's participation in any consortium is conditional on "data being made freely available," says Morgan. He adds that all the companies in the consortium are against intellectual property being taken on any data that it produces. ■

Novartis pins hopes for GM seeds on new marker system

Quirin Schiermeier, Munich

The multinational company Novartis Seeds last week launched a campaign to gain worldwide support from the plant-science community for its new marker-gene system, Positech.

The campaign follows the company's recent announcement that it plans to phase out antibiotic-resistance marker genes in its future products, in an attempt to restore public confidence in the safety of genetically modified foodstuffs.

Plant geneticists use marker genes to monitor the transformation of plant cells after a foreign gene has been introduced into a plant along with the marker. The most commonly used markers are antibiotic- or herbicide-resistance genes, which protect the cells against an agent introduced to kill those that have not taken up the desired DNA.

But concerns have been triggered about the potential threat to public health. There are fears that antibiotic resistance genes could 'jump' from transgenic plants to microorganisms such as gut bacteria, increasing antibiotic resistance in humans.

The Positech marker system uses transformed plant cells that contain the gene encoding the enzyme phosphomannose isomerase (PMI). In the selection process, only

the cells that take on the PMI trait can survive in a culture in which they are fed only mannose, a simple sugar. Novartis says the system has been tested successfully with transgenic maize and wheat.

"Positech is well along on the development path," Wally Beversdorf, head of research and development at Novartis Seeds, told last week's International Crop Sciences Congress in Hamburg. "We will have regulatory packages containing PMI probably within 12 to 18 months."

Basic researchers, who can use 'Positech' without royalties, have welcomed the launch. "The availability of less controversial marker genes would certainly improve the acceptance of our research, and cool down the public debate," says Uwe Sonnewald, head of molecular cell biology at the Institute for Plant Genetics and Crop Plant Research in Gatersleben, Germany.

But environmental groups say that phasing out antibiotic-resistance marker genes would only be a small improvement. Greenpeace has called Novartis' campaign "mere propaganda" aimed at easing consumer concerns about genetically modified foods. "Other important issues, such as the ecological impact of gene transfer into the environment, are ignored," says a spokesman. ■

Cambridge seeks £1.6 million to buy Newton's papers

Natasha Loder, London

An unrivalled archive of Sir Isaac Newton's scientific writings could be created if the University of Cambridge succeeds in its bid to buy a collection of documents and letters owned by the ninth Earl of Macclesfield.

Valued at £6.37 million (\$9.4 million), the collection is considered to be one of the most important in private hands at present. The university wants to unite the papers with its existing collection of related papers, which already forms the world's most complete collection of Newton material.

The UK Heritage Lottery Fund, which offers money from Britain's National Lottery, has already agreed to put £4.79 million towards the purchase to safeguard the heritage of objects important to "the formation of the character and identity of the United Kingdom".

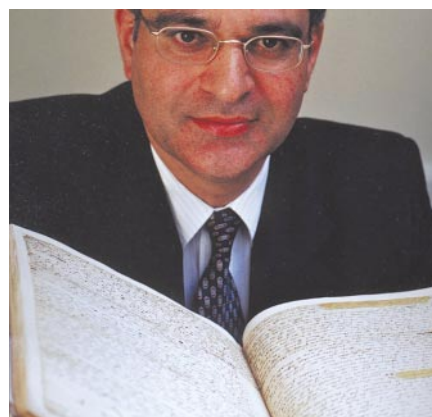
The university is now seeking the remaining funds from private benefactors. But Patrick Zutshi, keeper of manuscripts at the university library, says that even if it fails

to raise the money, the library would be prepared to go into debt to get the papers.

Newton, who lived from 1642 to 1727, spent nearly 40 years at the University of Cambridge, where he was appointed Lucasian professor of mathematics in 1669. It was during this time that he laid much of the foundations for modern mathematics, astronomy and physics.

The auction house Sotheby's is negotiating the sale of the papers on behalf of the Earl of Macclesfield. Although he is likely to have got a higher price through an auction, the Earl is keen for the collection to go to the university.

Other leading mathematicians of the day whose papers appear in the collection include Christiaan Huygens, Pierre de Fermat, Robert Boyle, John Flamsteed and Isaac Barrow. The collection also contains an unpublished 200-page manuscript *Astrologia Rationis Argumentis Solidis Explorata* by Nicolas Mercator, the astronomer and mathematician who invented the pendulum marine chronometer.



Zutshi: the Cambridge library is willing to go into debt to secure the Newton papers.

Peter Lipton, head of history and philosophy of science at the university, says if the purchase bid succeeds, conservation work will be carried out on the papers before they are put on public display. ■

♦ <http://www.lib.cam.ac.uk/>

♦ <http://www.sothebys.com/>

Embryo stem-cell work gets NIH go-ahead

Paul Smaglik, Washington

US researchers can at last apply for federal grants to study human embryonic stem cells. The go-ahead came last week when the National Institutes of Health (NIH) published the long-awaited terms under which it will fund such research. In principle, grants submitted now could be reviewed as soon as December and funded in the fiscal year 2001, which starts next October.

The move comes hard on the heels of a similar decision by the British government two weeks ago (see *Nature* 406, 815; 2000). It ends a de facto moratorium on US government support for embryonic stem-cell research, in place since the cells were first cultured successfully two years ago.

But the research allowed by the guidelines could be short-lived. The rules are based on an interpretation of the present laws banning the use of public funds for human embryo research, which could be reversed after this year's presidential elections.

In January 1999, the Department of Health and Human Services said that these laws — which are based on the fact that embryos are destroyed when the stem cells are extracted from them — do not apply to their stem cells, as these are not themselves embryos (see *Nature* 397, 185; 1999).

In a twist that even some stem-cell supporters find tenuous, the guidelines skirt the destruction issue by funding research only on the cells themselves. Extraction will be funded with private money and will use only surplus embryos that would be discarded after *in vitro* fertilization treatment. Both anti-abortionists and



Berg: it is too early to focus only on adult cells.

members of the National Bioethics Advisory Commission, which supports stem-cell research, have criticized the logic behind this.

The lack of a strong legal basis, together with opposition from the anti-abortion lobby, is making embryonic stem-cell research — already a volatile issue in the United States — a prominent political lightning rod. US presidential candidate George W. Bush has said through spokesmen that he opposes such research. If elected in November, some fear that he could sign an executive order forbidding it. In contrast, Democrat candidate Al Gore has said he supports stem-cell research.

The US Congress is also sharply split. Although some members have indicated that



Hot potato: opposition from anti-abortionists could still shift political views in Washington.

they would either sue the NIH to stop the research or draft legislation to prohibit it, others are trying to expand it. Senator Arlen Specter (Republican, Pennsylvania), for example, has drafted a bill that would allow federal funding for both use and derivation of embryonic stem cells.

Although the guidelines were issued in draft form last winter, the NIH has tried to defuse the controversy by creating a greater distinction between the derivation and the use of embryonic stem cells, and by applying tighter controls. For example, the new guidelines require, rather than recommend, an institutional review board's approval for the derivation of stem cells from both embryos and fetal tissue.

The most significant modification, says

Lana Skirboll, associate director for science policy at the NIH, is that the guidelines now allow donors of embryonic stem cells to be identified — the original guidelines required that their identity be masked.

Knowing the identity of donors could help minimize immune responses when cells are transplanted into other humans and also prevent infectious diseases from being transmitted.

Researchers will also have to specify what types of experiments they intend to perform. If they use stem cells for work not included by the guidelines, they could jeopardize other NIH funding for their institutions.

Scientific societies, including the American Society for Cell Biology and the Federation of American Societies for Experimental Biology have welcomed the guidelines, as have several patient advocacy groups. President Bill Clinton endorsed them, but Pope John Paul II denounced them.

Opponents say that the NIH should support more work on adult stem cells (see *Nature* 405, 6; 2000). David Prentice, professor of life sciences at Indiana State University and a member of Do No Harm, a group that opposes embryonic stem-cell research, says that adult stem cells have been unfairly maligned.

He says that recent research raises questions about the claims that adult stem cells cannot differentiate into as many cell types, or produce as many cells as embryonic ones. "They actually have a lot more promise and they don't have the ethical baggage," he says.

But Nobel laureate Paul Berg, professor of biochemistry at Stanford University, says concentrating only on adult stem cells would be premature. "This line of research is still in its infancy," he says.

Japan seeks science entrepreneurs

David Cyranoski, Tokyo

Japan's Ministry of International Trade and Industry (MITI) has linked up with the country's ministries of education and health to tackle a tricky problem: how to spur joint public-private research.

The result is the creation of two three-year grant programmes to fund joint research projects in new industries, with the hope of paving the way for commercialization and start-up companies.

Grants funded by MITI and the Ministry of Education, Science, Sports and Culture (Monbusho) will be targeted at information technology, energy/environment and biotechnology. Grants for work by industry and the Ministry of Health will target the

development of instruments for the diagnosis and treatment of heart disease and cancer, as laid out in the Medical Frontiers programme announced this spring.

The scale is small. But "the cooperation between ministries, and its establishment as a permanent part of the budget are quite significant accomplishments in a Japanese setting," says Yukio Kawaguchi, who runs MITI's collaboration with Monbusho.

MITI expects to receive most of the ¥1 billion (US\$9.3 million) it will request from the government for the grants with Monbusho, and probably a large fraction of its request for those with the health ministry. Selected projects will receive about ¥100 million from MITI; industry would

need to invest half that again.

A programme announced last year for joint research between industry and universities was financed from a supplementary budget, meaning that the money had to be spent in one year — a short period to develop new technologies, say MITI officials (see *Nature* 402, 116; 1999).

Koichi Sumikura of Tokyo University's Research Center for Advanced Science and Technology agrees that a three-year budget will "allow rollover and avoid some wastefulness of projects whose budget must be exhausted in one year". At the end of three years, government support would end and, it is envisioned, entrepreneurs would be able to exploit the joint research.

University researchers' enthusiasm for the project is still in doubt. The lifting of some restrictions, which allowed researchers to take up executive positions in companies (see *Nature* 403, 589; 2000), has yet to stir much interest.

Part of the problem, says Kawaguchi, is that "universities are less demanding than in the United States, and many researchers will tell you they are simply too comfortable to take a risk in industry". The current programme is designed to allow them to develop their technologies, and perhaps gain some new research equipment, without risk.

But problems with intellectual property could lessen industry's interest in joint research. Any action taken on a patent in Japan must be agreed to by all holders. Thus, even if a university were to hold only a small share of the patent, warns one Ministry of Health official, the prospective entrepreneur "would have to deal with lengthy bureaucratic deliberations on the use of the patent".

MITI's Medical and Welfare Equipment Industries section chief Yukiko Araki admits that, although MITI and the health ministry share the same goals, many details of funding and other procedural matters are still in need of attention. ■



Safe? Species such as this newly discovered Indonesian frog could benefit from the biodiversity plan.

Ecologists back blueprint to save biodiversity hotspots

Rex Dalton, Pasadena

Eminent ecologists have endorsed a multi-billion dollar blueprint for the conservation of biodiversity hotspots around the world, deflecting concerns about the scope and cost of the plan.

Conservation International (CI), a Washington-based non-profit organization invited leading lights of the biodiversity research community, such as Harvard University's Edward Wilson, to discuss the proposal at a four-day meeting at the California Institute of Technology last week.

"If we are going to have an impact on saving biodiversity, we have to approach this on a scale far beyond anything proposed before," says CI president Russell Mittermeier.

CI officials hope that philanthropists and private individuals will make the huge donations needed to support the conservation efforts. The first step towards this goal was taken recently, with the announcement of a \$75 million Critical Ecosystem Partnership Fund, which is expected to double in size within months (see *Nature* 406, 818; 2000).

The 25 biodiversity hotspots targeted for preservation under the blueprint cover 1.4% of the Earth's land area, but contain an estimated 44% of vascular plant species and 35% of all species in four major vertebrate groups. The hotspots — including areas of Latin America, Africa, Madagascar, India and Southeast Asia — were described earlier this year in an article by CI scientists and their co-workers (see *Nature* 403, 853–858; 2000).

Participants at last week's symposium said that conservation efforts should reach beyond these locations to save three wilderness regions threatened by forestry, mining and ranching. These are the Amazon Basin and the Guyana shield region in South

America, and the Congo Basin in Africa.

A draft document — called *An Agenda to Defy Nature's End* — was drawn up by the scientists to focus research programmes, develop local preservation efforts and reduce incentives for the destruction of biodiversity.

But the costs of preservation, management and running a research centre at each biodiversity hotspot could be nearly \$200 million a year in each case, CI officials say. Protecting the three wilderness areas could cost an additional \$4 billion, says Mittermeier, adding that he hopes to raise \$1 billion in private money to jump-start this effort.

CI's plan has sparked debate among ecologists, with some questioning whether there are enough data to identify the hotspots most worth saving. The meeting heard from some of the sceptics, including Andrew Dobson, an evolutionary biologist at Princeton University, and Andrew Balmford, a zoologist at Cambridge University, who have publicly challenged the proposal (see *Nature* 405, 393; 2000).

Stuart Pimm, a conservation biologist at Columbia University in New York who led the discussion of the blueprint, says that "a broad consensus on what needs to be done" emerged from the meeting.

But Jeff McNeely, chief scientist for the Swiss-based World Conservation Union says that CI needs to reach out to governments whose nations include the hotspots, so as to introduce more reality into its planning.

"Governments often don't like biodiversity," McNeely points out. "They are trying to build nations. I am very supportive of what CI is trying to accomplish. But if you are going to design something that works, you have to talk to the opposition." ■

♦ <http://www.conservation.org>

♦ <http://www.defyingnaturesend.org>

North Carolina reflects on ammonia controls

Jessa Netting, Washington

State legislators and environmental officials around the United States are watching their colleagues in North Carolina closely to see what action they take to control ammonia emissions — currently unregulated — from the state's animal production facilities.

Next week, members of a scientific committee, assembled by state agencies in North Carolina to investigate ways of processing and storing the waste, will meet to decide whether to recommend immediate regulation to meet concerns that the emissions are a potential health and environmental hazard.

Some members of the committee — including its chair, Wayne Robarge, a soil physical chemist at North Carolina State University — argue that more scientific and technical assessments are needed to ensure that its recommendations are well informed.

Pai Yei Whung of the National Oceanic and Atmospheric Administration also says that more study is necessary. But she points out that, as there is currently no federal rule regulating ammonia concentrations at levels that are not immediately hazardous to health, the large animal production industry can claim "that they aren't violating any laws".

Gaseous ammonia is washed out of the



In the dock: pig production has been blamed for rising ammonia levels — but are they a hazard?

atmosphere when it rains, and enters streams, rivers and estuaries in North Carolina, where it increases the amount of nitrogen in the system. If it is converted while airborne to ammonium nitrate, it could make its way even further afield, to the estuaries and bays along the Maryland coast.

The Environmental Protection Agency has recently targeted ammonium nitrate for regulation under Clean Air Act standards on the grounds that it is a danger to human health. And gaseous ammonia may account for up to half of the nitrogen levels in some waterways along the Atlantic coast, Robarge and other chemists told a meeting of the

American Chemical Society in Washington last week.

But scientific uncertainty surrounds the relationship between how much gaseous ammonia emitted will rain into bodies of water, and how far ammonia and its products can be transported by the wind. This has prompted Robarge to oppose regulation until the impacts and influence are clear and the alternative technologies have been sorted out.

From a purely air-quality perspective, Whung describes levels of ammonia immediately surrounding swine-production operations as "disturbing". She has monitored ammonia in North Carolina with Robarge, and is now concerned with atmospheric ammonia deposition in Chesapeake Bay, Maryland, which is within North Carolina's 'air shed'.

At last week's American Chemical Society meeting, evidence was presented that cars, rather than livestock, may be the main source of ammonia in the atmosphere. In a study of 4,500 vehicles, researchers from the Oak Crest Institute of Science in California found ammonia levels from car exhausts were so high that they estimated cars are adding twice as much ammonia to the air of the state's southern coastal basin as livestock. ■

Genome website set up to help with sequence analysis

London After PubMed Central and Biomed Central comes 'Human Genome Central' — a 'master website' being set up by the International Human Genome Sequencing Consortium to provide access to tools for analysing the sequence data being produced by its members.

The consortium says that although the raw sequence data are already available in three public databases — Genbank, the European Molecular Biology Laboratory Database and the DNA Database of Japan — analysing this data is a daunting and time-consuming task.

The new website has therefore been designed to give users "comprehensive and comprehensible" ancillary information and tools. These should provide visitors to the site with a picture of the genome that is continually updated.

The information available will include the overlaps between clones, the location of each clone, an integrated sequence that merges the individual clones, and annotation of gene content.

♦ <http://www.ncbi.nlm.nih.gov/genome/central>

♦ <http://www.ensembl.org/genome/central>

Japanese doctors took brain samples without consent

Tokyo Two doctors in the Tokyo medical examiner's office, who were revealed last week to have taken brain samples during autopsies without permission from the office or the families of the deceased, have pleaded ignorance of the need for consent. Since 1992, the two have taken nearly 100 such samples from people who had died of nervous diseases, passing them on to researchers at the Tokyo Institute of Psychiatry.

A Ministry of Health official said that there were probably more undiscovered cases. But he added that the incident did not reflect any ambiguity in the regulations, and should not affect future research as it had not stirred up "the kind of public outcry and legal action that might happen in the United States".

'Healthy' cattle might still harbour BSE

London Scientists from the UK Medical Research Council's prion unit have found evidence of a 'sub-clinical' form of BSE in mice, raising the possibility that apparently healthy cattle could harbour the disease. The researchers also found that prions transferred easily from hamsters to mice, a species barrier that was thought to be robust.

According to the Department of Health, the results, published in *Proceedings of the National Academy of Sciences*, will be considered by the Spongiform Encephalopathy Advisory Committee when it meets at the end of September. "Current measures to protect public health were introduced on the basis that infection in animals and people may be present in the absence of clinical disease," the ministry said in a statement.

US climate change report comes under fire

Washington Fifteen industry groups have joined forces to label an assessment of the potential consequences of climate change for the United States a "failure". They were especially critical of the initiative's recent report *Climate Change Impacts on the United States*, which discusses how global warming could affect different parts of the country (see *Nature* 405, 725; 2000).

The groups claim that the predictions play down the amount of uncertainty involved. The National Association of Manufacturers, American Farm Bureau Federation, American Petroleum Institute, US Chamber of Commerce and the National Mining Association are among those in the campaign.

Chasing the dragons

Stunning fossils from Liaoning province have created a boom for Chinese palaeontologists and local farmers alike. Rex Dalton reports from the wild frontier where researchers do battle with the black market.

The farmers of Songzhangzi, some 300 kilometres northeast of Beijing, had always ignored the rocky ridges that loom over their small village. Then they struck palaeontological gold. About three years ago, realizing that some people will pay good money for fossils, the villagers took out their picks and shovels and began burrowing into the strata.

The hillsides were soon honeycombed with grave-sized shafts, and fabulous specimens began to emerge. Those that fell into the hands of palaeontologists include *Liaoxiornis delicatus*, the smallest known bird from the Mesozoic¹, and *Hyphalosaurus sinohydrosaurus*, a long-necked diapsid reptile². More await description in the literature.

But others have almost certainly been lost to an illicit trade, as impoverished locals sold their finds to the highest bidder. And although the Institute of Vertebrate Paleontology and Paleoanthropology (IVPP) in Beijing brokered a deal with provincial officials to dig at the site, known as Dawangzhangzi, earlier this year, Songzhangzi today resembles a gold-rush frontier town, where science is losing out to commerce.

The story has been repeated at other important sites across Liaoning province, located between Beijing and the North Korean border. Chinese palaeontologists dig side-by-side with local farmers, knowing that stunning specimens are being lost to science

all the time. In late May, some 30 leading palaeontologists from around the world — and me — experienced this difficult working environment for themselves.

Escorted by scientists from the IVPP, we visited quarries, museums and roadside digs on a four-day tour of Liaoning, organized as a prelude to the fifth quadrennial meeting of the Society of Avian Paleontology and Evolution (SAPE) in Beijing. The western researchers were awed by the breathtaking quality and evolutionary significance of the fossils on display. But we also learned just how easily important specimens can vanish into the black market for sale to private collectors.

The Liaoning beds have yielded a host of primitive birds and the feathered theropod dinosaurs that most palaeontologists believe are their closest relatives^{3–11}. Together, these fossils have revolutionized our understand-

ing of bird evolution, and made the Chinese scientists behind the discoveries among the most cited in their field. But it is not simply the good fortune of discovering these fossils that has propelled China's palaeontologists to the fore.

As in many disciplines, China has sent bright young researchers to be trained in the West's top institutes. Among them is Zhou Zhonghe, who completed his PhD at the University of Kansas in Lawrence last year. But the palaeontological buzz in China is now so great that he rushed back home, rather than stay abroad as a postdoc as Chinese scientists in other disciplines often do. "You have this chance to make these kinds of discoveries but once or twice in a century," says Zhou. "It is a great feeling, and the discoveries are not slowing down."

But these growing East–West links have brought tension, mainly due to restrictions placed on access to the Chinese fossils. Several young American palaeontologists attending the SAPE meeting were disappointed when they were only allowed to study a few specimens, despite repeated pleas to see more. And some Chinese scientists studying in the United States have not shared specimens as their collaborators had expected.

Buried treasures

But there were no complaints when palaeontologists on the SAPE tour witnessed the unearthing of a specimen of *Confuciusornis sanctus*, one of Liaoning's famed early bird fossils, at a dig near Sihetun, east of the city of Chaoyang. IVPP scientists have mined this hilltop site for several years. The excavation goes to a stratum dated to about 124 million years old, in which a volcanic eruption left avian specimens entombed on an ancient lake bed.

Workers from the local village used putty knives to begin scraping away the thin ash layer in a flat section about five metres square. Halfway through the section, the *Confuciusornis* was unearthed. "We are lucky to be alive when this material is being discovered," says Rick Prum, an evolutionary biologist at the University of Kansas. The excavators themselves were less effusive, however. One of them, Cai Yizhong, was asked his feelings about the fossil. "No feelings," he replied, via a translator.

Indeed, to many locals, interest in the fossils begins and ends with their cash value.

A fossil dealer had set up a makeshift stall in the lobby of the palaeontologists' hotel.



Reeling them in: fossils of small fish unearthed at Songzhangzi.



Catching the early bird: locals at Sihetun work to reveal what appears to be a *Confuciusornis* (right) to visiting western palaeontologists. But such fossils often end up on the black market.

Such specimens sell for thousands of dollars on the international market, and the villagers of Liaoning province know it. Before the SAPE group arrived, the Sihetun quarry had been filled in with loose rocks to prevent pilfering. It took weeks for villagers with wheelbarrows to remove this protective fill before the IVPP could conduct its demonstration.

Despite such measures, the fleet of new motorcycles driven by the village's young men — who make at most US\$3 a day digging for the IVPP — pointed to earlier successful sales of fossils. Villagers get only a fraction of the international black market price, but even this is a massive improvement over what they normally earn in the local agrarian economy.

There are similar sites throughout the region. For instance, just outside the burgeoning city of Jinzhou, alongside a country road near to a village, scientists from a local museum have discovered a deposit from the early Cretaceous. Erosion had already done much of the excavation: fish fossils were found without digging, and avian specimens are believed to lie just a metre or two below the surface.

Hot rocks

Preventing the pillaging of this site near a city of several million people will be a challenge. The legal issues are confused, even to the scientists involved. "It is very complex," says Zhu Min, IVPP's scientific director, who notes that the Chinese central government is in the process of rewriting laws on fossil collection. "This issue is very hard to deal with."



Officially, the government considers the fossils — in particular the avian specimens, dinosaur eggs and certain mammalian skeletons — as national treasures that cannot be sold legally. But privately, Chinese palaeontologists say that Beijing exerts little control over what happens in the countryside. Provincial officials, meanwhile, are clamouring for more autonomy, which can cause problems for Beijing-based palaeontologists (see "The world's loneliest museum", right). In this transitional period corruption has become rife in provinces such as Liaoning, and controls over the illicit fossil trade are lax, to say the least.

Western palaeontologists on the SAPE tour did not have to look far to learn about the problems. They were right in front of us in the lobby of our hotel in Chaoyang: a fossil dealer had set up a makeshift stall. This offered a variety of commonly sold fish and reptile fossils. But it also offered specimens of

The world's loneliest museum

As well as the illegal fossil trade, palaeontologists from Beijing must also deal with the complexities of provincial politics. Relations with officials in Liaoning can be strained, as western palaeontologists learned when they visited the new Beipiao Fossil Museum, 12 kilometres outside the city of the same name.

The museum houses an enviable collection of fossils. A stunningly preserved *Caudipteryx* — one of only six known specimens of this feathered creature — takes pride of place. One palaeontologist remarked that any museum in the world would pay a hefty price for such a specimen.

Despite its palaeontological riches, the Beipiao Fossil Museum is one of the most inaccessible anywhere. The gravel road leading to it is so bad that buses have to stop about a kilometre away. The museum cost about US\$1.2 million, according to Beipiao officials, the cost being split between local and central government. Proudly displayed photos show Zou Jiahua, a top official of the Chinese National People's Congress, dedicating the museum's construction.

But nobody seemed to know when it would open to the public, and the entire project had Beijing-based palaeontologists shaking their heads. "I have no idea why they put it there," said one. "But we can say nothing. If the locals in Beipiao are against you, you can do no collecting there."

The museum is also draining funds from the Institute for Vertebrate Paleontology and Paleoanthropology in Beijing. The valuable specimens require full-time guards, and Beipiao officials are charging the cost of the guard squad to the institute. They also double-charged the institute for a lunch given to visiting western palaeontologists.

These expenses can have a real impact on the Beijing institute, which has such a limited budget that it was unable to repair the elevator and some toilets in time for the June meeting of the Society of Avian Paleontology and Evolution, held at its six-year-old museum.

Confuciusornis, trade in which is supposed to be strictly controlled, if not banned outright.

Other dealers were more subtle. One evening in Chaoyang, operators of a small, private 'museum' offered to show interested scientists their specimens. Half a dozen of us jumped into cabs. We rode through this gritty city of three million people, across rail tracks that still carry steam engines, and into

► a poor neighbourhood. Away from the main boulevards, down an alley in a walled courtyard of a private residence, we found the 'Long Chen Fossil Museum'.

A garage-type building had been converted into a showroom, secured by steel doors. Inside, there was a *Confuciusornis*, fossils of mammals, fish and reptiles along with petrified wood. Some items carried price tags. The 'museum' operators sported the latest cellular telephones and drove a luxury new four-wheel-drive vehicle — a stark contrast in a neighbourhood where bicycles are the norm. They responded to jokes in English, but wouldn't answer my questions about their activities. A Chinese researcher said that he was once asked to facilitate the museum's sales. And the lobby salesman tried to sell a specimen to a visiting palaeontologist. Both turned down the offers.

Occasionally, there are arrests for fossil dealing. One person was detained last year at the port of Dalian trying to smuggle a dinosaur egg out in their baggage; an illicit shipment of fossils was discovered earlier this year at Beijing airport; and about two weeks before the SAPE meeting, police raided the Panjiyuan flea market in Beijing, seizing fossils and dinosaur eggs.

But given the open trade of fossils in Liaoning, and the availability of Chinese specimens on the western black market, enforcement seems more the exception than the rule. Last February, at a big fossil show in Tucson, Arizona, one dealer of Chinese origin, now based in the Los Angeles area, had dozens of *Confuciusornis* specimens stacked like newspapers. Dealers often claim to have documents from Chinese institutions authenticating the scientific exchange of such specimens. But from what I saw in Liaoning, these records can be created as easily as someone can stock specimens in a 'museum' in a back alley.

Chinese palaeontologists wage a constant battle to prevent the fossils ending up on the black market. Another disturbing aspect of this trade is that fossils can be damaged by forgers, who amalgamate pieces from different specimens in the hope of creating a fossil that will attract a higher price.

The prevalence of forgeries was made clear on the first day of the SAPE field trip. Word of an exciting new specimen at a museum in Jinzhou had spread to Beijing. But when the fossil was displayed on a table in the museum lobby, scientists from the IVPP and SAPE quickly realized it was a composite of several specimens. The forgery was so obvious that even an amateur could see two tibiae of different diameters were pieced together end-to-end to create a 'leg'.

Keep it real

IVPP scientists are working to educate both local farmers and provincial museum curators not to dress up fossils, emphasizing they are more valuable in their natural state. But it is an arduous process. Xu Xing, an IVPP graduate student who has worked for years in Liaoning province, says that there are "assembly line factories" where workers piece together fossils for sale on the black market. Earlier this year, he visited one of these factories to try to deter the practice. Xu hopes his message hit home, but he is worried that similar workshops will set up, given the black market's demand for fossils.

'*Archaeoraptor liaoningensis*' is the most famous forgery. Stephen Czerkas, the operator of a small Utah museum, had bought the specimen in the belief that it was an important link in bird and dinosaur evolution, and the fossil was described in the November 1999 issue of *National Geographic* magazine.

Archaeoraptor seemed to be a bird, but it had a tail similar to a group of theropod



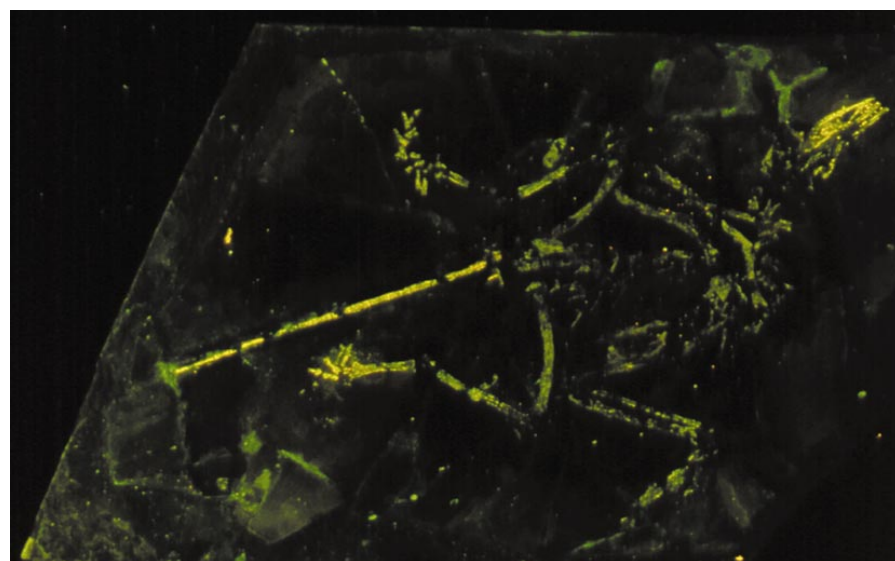
Home again: Xu Xing returns from the United States with *Archaeoraptor*.

dinosaurs called dromaeosaurs. A panel of palaeontologists easily debunked the specimen earlier this year, with a computerized tomography scan confirming the specimen as a composite¹².

Czerkas repatriated the fossil to the IVPP in May. Xu, meanwhile, ventured into the netherworld of Chinese dealers last December in search of the rest of the theropod from which *Archaeoraptor*'s tail came. After locating it, he bought the remains from a dealer for \$5,000, supplied by the National Geographic Society, which is now seeking to uncover *Archaeoraptor*'s true origins.

But since then, local dealers have learned that Czerkas paid \$80,000 for *Archaeoraptor*. Now they are hiking up their prices, making it harder for the IVPP and other Chinese research institutes to buy specimens and rescue them from the illegal trade. As Xu and his colleagues work on manuscripts to describe both new species that made up *Archaeoraptor*, they can only wonder about the other stunning specimens that are disappearing into private collections without ever being scientifically described.

Rex Dalton is Nature's West Coast US Correspondent.



Famous forgery: *Archaeoraptor liaoningensis* turned out to be a composite of two fossils.

- Hou, L., & Chen, P. *Chin. Sci. Bull.* **44**, 834–838 (1999).
- Li, J., Zhang, B. & Li, Q. *Mem. Beijing Nat. Hist. Mus.* **56** (suppl.), 1–7 (1999).
- Hou, L., Zhou, Z., Gu, Y. & Zhang, Z. *Chin. Sci. Bull.* **40**, 1545–1551 (1995).
- Hou, L., Zhou, Z., Martin, L. D. & Feduccia, A. *Nature* **377**, 616–618 (1995).
- Hou, L., Martin, L. D., Zhang, J.-Y. & Feduccia, A. *Science* **274**, 1164–1167 (1996).
- Ji, Q. & Ji, S. A. *Chin. Geol.* **233**, 30–33 (1996).
- Ji, Q. & Ji, S. A. *Chin. Geol.* **238**, 38–41 (1997).
- Chen, P., Dong, Z. & Zhen, S. *Nature* **391**, 147–152 (1998).
- Qiang, J., Currie, P. J., Norell, M. A. & Ji, S. *Nature* **393**, 753–761 (1998).
- Xu, X., Tang, Z., & Wang, X. *Nature* **399**, 350–354 (1999).
- Xu, X., Wang, X. & Wu, X. *Nature* **401**, 262–266 (1999).
- Dalton, R. *Nature* **403**, 689–690 (2000).

Sequencing solution: use volunteer annotators organized via Internet

Sir—When the *Pseudomonas aeruginosa* genome project began in 1997, one question facing us was how best to annotate gene descriptions and other information. We decided to take a community approach, in an attempt to improve the quality of annotation and to use the resources of all those researchers working with this versatile pathogenic bacterium.

Our results support aspects of 'open annotation' approaches for the human genome, as described in Correspondence¹ and News². However, our experience has suggested that certain precautions must be taken.

For the community project, termed PseudoCAP³, we recruited volunteers from the *Pseudomonas* research community, and later others, to submit annotations of genes or gene families with which they were familiar, through the direction of a single project moderator. Unlike the annotation jamboree for the *Drosophila* genome project⁴, all communications with, and submissions by, the volunteer participants were made exclusively through the Internet⁴.

The PseudoCAP annotations were overlaid on a genome viewer console developed by PathoGenesis, containing layers of other automatically generated analyses and literature reference information.

This resource, coupled with a critical, conservative annotation approach, was used to generate the final genome annotations, which were also classified according to whether they were based on (1) functional studies in *P. aeruginosa*; (2) high homology to functionally studied genes in other organisms; (3) low homology to functionally studied genes; or (4) homology to hypothetical genes (see the accompanying paper in this issue⁵).

We were pleasantly surprised at the enthusiasm for PseudoCAP—61 participants made 1,741 submissions. Most of the later participants did not work on *Pseudomonas* but were researchers who wanted to examine genes of particular function. Judging from this response, an adequate number of annotators could probably be recruited for other community annotation projects.

Given the experimental nature of our approach, we allowed participants to submit whatever information they wished; as a result, variation in the quality of annotations led to numerous inconsistencies. Therefore, review of all annotations by a core group was essential. For the future, we recommend that community participants

should be required to clearly define their annotation methods and criteria for using any particular functional description, and adopt a consistent, searchable format. Otherwise inconsistencies will not be easily detected, and useful information (for example, retrieval of all annotations based on a certain type of functional study) will not be readily available.

Final annotations for the genome project were based almost exclusively on functional studies of the gene in question, or on close homology of the encoded protein to functionally studied proteins. This method involved significant manual intervention, which could be automated if a sequence database based only on functional studies is created. SwissProt and the National Center for Biotechnology Information's RefSeq are beginning to develop in part along these lines.

In the meantime, we recommend that genome projects consider a community-aided annotation approach, coupled with critical, conservative annotation by a core group of project annotators. If such community involvement occurs through the Internet in a formal, well publicized setting, annotations can continue to be updated and corrected after a genome sequence is published.

Fiona S. L. Brinkman*, Robert E. W. Hancock*, C. Kendal Stover†

*Department of Microbiology and Immunology, University of British Columbia, Vancouver, British Columbia V6T 1Z3, Canada

†PathoGenesis Corporation, 201 Elliott Ave West, Seattle, Washington 98119, USA

1. Hubbard, T. & Birney, E. *Nature* **403**, 825 (2000).
2. Butler, D. *Nature* **404**, 694 (2000).
3. <http://www.cmdr.ubc.ca/bobh/PAAP.htm>
4. Pennisi, E. *Science* **287**, 2182–2184 (2000).
5. Stover, C. K. *et al. Nature* **406**, 959–964 (2000).

Let's get the right man in the right job

Sir—As the late Peter Medawar¹ wrote "Scientists are entitled to be proud of their accomplishments, and what accomplishments can they call 'theirs' except the things they have done or thought of first?"

In this spirit I would like to correct the recent News and Views article² on the second Chapman Conference on Gaia. In an otherwise interesting article, Jim Gillon gets both my name and affiliation wrong, calling me D. Williamson of the University of Liverpool rather than D. Wilkinson of Liverpool John Moores University.

It may be of interest that some of my ideas on Gaia and evolution were published in *Oikos* last year³; since writing that paper I have become slightly less sceptical about the possibility of global regulation, for the

reasons summarized in Gillon's article.

David M. Wilkinson

*Biological and Earth Sciences,
Liverpool John Moores University, Byrom Street,
Liverpool L3 3AF, UK*

1. Medawar, P. *Pluto's Republic* (Oxford Univ. Press, Oxford, 1982).
2. Gillon, J. *Nature* **406**, 685–686 (2000).
3. Wilkinson, D. M. *Oikos* **84**, 533–553 (1999).

Model already exists for fair use of gene data

Sir—I agree wholeheartedly with the analogies drawn between the advantages of open source and the genome projects by Russ, Aparicio and Carlton¹. However, I would go further than Alberts and Klug^{2,3} and suggest that it should never be possible to patent a gene. A gene is a pre-existing entity; it can be discovered, but not invented. Of course a drug invented to exploit a gene, or a method using a particular gene for therapy, is a different matter. These are clearly inventions.

I should also like to clear up the details of the GNU Public Licence. The GNU licence is not about software being free of charge. It is about freedom: allowing the user of the code to do with it what they will. The licence actually comes in two forms: the GPL and the LGPL. The GPL allows use of the software in commercial or freely distributable software. The restriction, however, is that the distributor of the software must make the source code available and must pass on the same rights of freedom; any code linked into an executable with GPL code must also fall under the GPL. Thus a product may be sold including GPL code, but the purchaser has the right to distribute the product without restriction, either for a fee or at no cost.

The LGPL, on the other hand, allows freedom of distribution (either commercially or at no cost), but allows executable programs to be created which do not require the LGPL code to be distributed and can be sold with the usual commercial restrictions. Thus the genome model better fits the LGPL. The original gene data should be completely free and shared throughout the community. Companies would then be free to exploit that information and to create patentable commercial products based on it. More than one company would be able to exploit the same gene, putting an end to the current 'land grab'⁴.

Andrew C. R. Martin

*School of Animal and Microbial Sciences,
University of Reading, PO Box 228, Whiteknights,
Reading RG6 6AJ, UK*

1. Russ, A. P., Aparicio, S. A. J. R. & Carlton, M. B. L. *Nature* **404**, 809 (2000).
2. Alberts, B. & Klug, A. *Nature* **404**, 325 (2000).
3. Alberts, B. & Klug, A. *Nature* **404**, 542 (2000).
4. Gavanagh, H. *Nature* **404**, 684–686 (2000).

The tree of ignorance

Evolution emerges unscathed from the battle of creation.

The Triumph of Evolution and the Failure of Creationism

by Niles Eldredge

Macmillan: 2000. 223 pp. £16.95 & \$24.95

Robert W. Cahn

In the third century BC, the Greek philosopher Aristarchus conceived the heliocentric theory of the heavens. For his pains, Cleanthes, the head of the Stoics, a militant religious group, sought to have him indicted for impiety. Two millennia later, the Catholic Church did indict Galileo for the same impiety. In 1925, in Tennessee, a group of religious fundamentalists, the 'creationists', who insisted (as they still do) on the literal truth of the Old Testament account in Genesis, sought and failed to use legal recourse to prevent a schoolteacher from teaching evolution. *Plus ça change, plus c'est la même chose.*

Niles Eldredge, a palaeontologist, has become infuriated by the antics of modern-day creationists, who still strive to use political activism and legal action to curb the teaching of evolution in US schools. This means that they have to get around the massive evidence in support of evolution. For instance, they insist on the 'gospel truth' of a bishop's calculation that the Earth is only some 4,000 years old. Accordingly, it must follow that all conclusions based on fossils and their dating are an illusion. God put the fossils into the Earth to confound impious followers of the archdevil, Darwin, or falsified their geological sequence.

Eldredge's book is in part a very clear and informative thumbnail sketch of the history of evolutionary theory, and of the experimental findings on which it is based. He places special emphasis on the uneven pace of change, stasis alternating with rapid changes at times of great evolutionary stress. He also analyses the ever-growing links between evolutionary theory and genetic insights.

Furthermore, he offers a hair-raising account of the creationists' twists and turns, in particular their repeated attempts to change State law so that biology teachers are forced to give 'creation science' equal time and emphasis as evolution. Although they seek scientific respectability, Eldredge is at pains to make it clear that creationists are not scientists. A point he stresses particularly is that creationists regard disagreements on points of detail between evolutionists as evidence against their scientific credibility.

Creationist scientists, we are told, only accept as 'facts' what all creationists accept as a matter of faith. In fact, such 'facts' are by their nature unfalsifiable, and therefore not scientific facts at all.



Creative thought: can evolution be beaten by a question of faith?

Eldredge tries valiantly to counter creationism with scientific arguments, as befits a respectable working scientist. He mostly resists the temptation to question creationists' motives; after all, a proper scientist does not ask whether an opponent's mother failed to nurture him properly or his father was unduly authoritarian. The one exception to this self-denial is the repeated remark that creationists are convinced that the (Christian subset of) humanity will behave morally only if they submit their intellects in every particular to the word of God as revealed in the Old Testament. It is surprising that those who call themselves creationist scientists do not call to mind the injunction to eschew the fruit of the Tree of Knowledge.

The book does not go into the distinction between mythos and logos as expounded in Karen Armstrong's splendid book on religious fundamentalism (in several faiths), *The Battle for God* (HarperCollins, 2000). Genesis is mythos, evolution is logos,

and as she says, "the two are essentially distinct, and it was held to be dangerous to confuse mythical and rational discourse". That is precisely what creationists do.

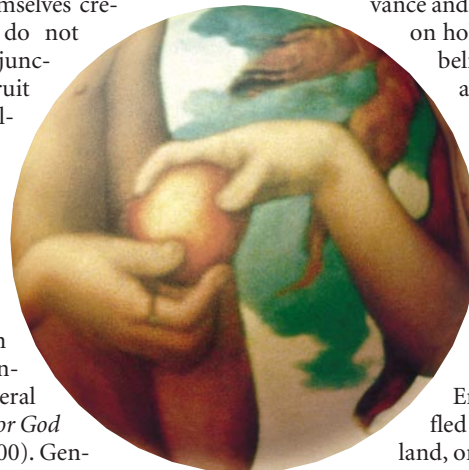
This leads me to suggest that perhaps the only way to counter the steady menace that creationists represent to good scientific education is to question their motives as publicly as possible; that is, to do what good scientists are trained not to do. In my capacity as a non-American, non-biological and non-Christian scientist, I want to raise questions of a wholly non-scientific nature. Why have creationists emerged as a political force (for that is what they are) only in the United States? Further, why has creationism only surfaced among Christians, and not Jews and Muslims (for whom the Old Testament is also a holy book), even though these faiths include their fair share of extreme fundamentalists?

Perhaps it is the US constitution, and its emphasis on States' rights, that has given US creationists a tool ideally suited to their purpose. It seems that, time and again, their claims for 'equal time' have been so cunningly reformulated that even Federal Supreme Court rulings can be bypassed.

For an outsider such as myself, educated when T. D. Lysenko in the Ukraine was, with Stalin's support, destroying Soviet genetics by enforcing a Lamarckian orthodoxy, attempts to lay down scientific correctness using the law are immensely dangerous. Perhaps evolutionists should give up trying to counter creationists by convincing them scientifically (a hopeless task), and simply appeal to the public and to judges in terms of clear and present political danger.

But what intrigues me is: why is creationism confined to US protestant fundamentalists? Ultraorthodox Jews and Muslims place more emphasis on ritual, dietary observance and behaving appropriately

on holy days, than on correct belief. They would perhaps approve Elizabeth I's assertion that she was not inclined to open windows into men's souls. By contrast, creationists seem to be wholly focused on belief. It would take a better historian than me to assess how far the legacy of the New England puritans who fled intolerance in Old England, only to establish an intolerant society in the New World, has



book reviews

worked through to modern creationism, and if that is indeed the prime source of the scourge, why it migrated from New England (where creationism is not today a menace) to the South.

Meanwhile, Eldredge's book, readily accessible to any educated reader, deserves to be widely read. ■

Robert W. Cahn is in the Department of Materials Science and Metallurgy, University of Cambridge, Pembroke Street, Cambridge CB2 3QZ, UK.

Martian chronicles

In Search of Life on Mars

by Malcolm Walter

Perseus/Allen & Unwin: 1999. 170 pp. \$25, £8.99

Dead Mars, Dying Earth

by John E. Brandenburg and Monica Rix Paxson

Element: 1999. 306 pp. £16.99, \$26.95

Kenneth Nealson

The search for extraterrestrial life has traditionally focused on Mars, often with acrimonious results. The 'pro-life' groups claim that unquestionably there is, or has been, life on Mars. The 'pro-choice' groups take a more cautious view, often paraphrasing Carl Sagan ("Extraordinary claims require extraordinary results") and Richard Feynman ("Unfortunately, you are the easiest person [for you] to fool"). The acrimony often arises because the pro-lifers view such caution as an anti-life stance, while the pro-choice group would gladly be convinced by high-quality data.

In these two books, we have some classic examples of these views. Malcolm Walter explains the difficulties of ascertaining with certainty the nature of life, even on this planet, and suggests some cautious approaches to getting the right answer. Brandenburg and Paxson, on the other hand, accept that life has been discovered on Mars and elsewhere, and wonder (often loudly) why the rest of the world cannot agree. It is a classic case of the believers losing their scientific perspective — seduced by their own ideas — and the questioners being accused of irrationality.

How can the search for life on Mars be framed and focused by Earthly studies? This interesting question has been addressed by palaeontologist Malcolm Walter in his short, readable treatise. Walter has worked for more than 35 years on fossils from our planet, focusing mainly on the ancient stromatolites of Western Australia. When such fossils are properly collected and analysed they sometimes reveal 'footprints in the sand' of past life on our planet. More often they yield difficult-to-interpret data that do not prove the existence of life. With this background, Walter discusses the frustrations that we now



Down to Earth

Building Planet Earth: Five Billion Years of Earth History (Cambridge University Press, \$39.95, £25.00) by Peter Cattermole details the geological evolution of Earth. Through a range of colourful diagrams and photographs the book explains the mechanics behind the planet's development. Sampling fossils and

lava on the way, the book reveals the origins of the continents and discusses what causes modern-day natural disasters. Familiar features, such as Ayers Rock (above), are described and their genesis detailed. The book also looks at the methods used in geophysics and geochemistry.

face with the current Mars meteorite analyses, and perhaps will face with returned Mars samples.

Walter is pro-choice: he has a healthy scepticism about the existence of life on Mars and spends a good deal of time discussing why one should have such a view. Yet he maintains an open mind. In particular, there are discussions of the problems encountered in dealing with samples that have sat on our planet for many years. Contamination by both organic materials and living organisms is virtually impossible to prevent. These problems, coupled with the use of structures to infer life (and the ability of many mineral forms to mimic structures), make the unambiguous interpretation of structural data difficult. As Walter points out, the return of pristine samples from Mars would be a huge step forward, although negative results from a few sites will not prove that life does not exist, or has never existed, on Mars.

A key to this book, and a major difference between it and the following book, is Walter's appreciation of the prokaryotes (microbial cells and ecosystems). He makes strong arguments as to why searching for microbes should be part of the strategy for life-detection, how one might do it, and how this relates to studies of ancient Earth. To Walter, finding microbial life would be as significant as finding any other kind, and the nature of that life, if found, would have major impacts on many facets of Earthly life.

Throughout, Walter expresses his feelings about the joy of science, the thrill of the hunt, the difficulty of being sure about any-

thing in the complex geobiological setting, and the frustrations of being half-right, or even wrong, as the work has proceeded.

In almost every way imaginable *Dead Mars, Dying Earth* represents a stark contrast to Walter's book. I found it difficult to read: it moves from subject to subject distractingly. It is actually two books in one — the first a discussion of the history of Mars and the arguments for and against life there. The pro-life arguments are readily accepted, while those against are largely dismissed. The authors make a number of claims without showing any evidence to support them. By the end they have 'established' the history of Mars, the existence of life on the planet, how it was discovered, how and when it was extinguished, and how this relates to the present global-climate situation on Earth today.

The second part is a discussion of the global CO₂ and oxygen problem, and an attempt to suggest some ways out of the dilemma. This was much more palatable. I could not agree more with the authors about the importance of the problem, the apparent apathy of humans towards it, and some of the suggestions for improving the situation. There is overwhelming evidence of the seriousness of this issue, some of which is presented by the authors.

The authors' view of life is interestingly eukaryote-centric. Their entire discussion ignores any microbial life that might escape UV radiation by being a few metres underwater, and the potential role of such non-oxygen-breathing life in maintaining the planetary balance. I would have appreciated

some nod to the microbes in recognition of their ability to produce oxygen, consume CO₂ and generate other gases, but the authors ignore them both on Earth and on Mars. Maybe what is being discussed here is just another species extinction (this time, humans) on a planet that will do perfectly well without them.

The book is a curious attempt to discuss two major subjects that appear to have at best a tenuous connection. It ends with the authors sitting on Mars, breathing the oxygen-rich atmosphere that has been created, and discussing how good it is that they have saved the Earth. I appreciated this ending, as it put the entire treatise in a fictional mode; I was now free to take it all less seriously. However, I doubt that this is what the authors intended. ■

Kenneth Neelson is in the Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91125, USA.

Bad eggs

Cuckoos, Cowbirds and Other Cheats

by N. B. Davies

Academic: 2000. 310 pp, \$29.95, £24.95

Robert C. Fleischer

To many people, even some who profess a deep love of nature, brood-parasitic birds such as cuckoos and cowbirds are evil creatures that do not deserve to exist, merely because they deceptively manipulate songbirds into incubating their eggs and rearing their young. For example, William Dawson, in his 1923 book *The Birds of California*, poetically describes the cowbird as “a blight upon the flower of Progress” and “the unchaste mother of a race gone wrong”.

There is apparently something extremely distasteful about birds that forgo their parental responsibilities and dump their offspring in the nests of innocent strangers. However, to me, and obviously to Nick Davies, the author of this wonderful new book on brood parasitism in birds, the 100 species of brood-parasitic birds provide some of the most fascinating and intricate examples of the workings of natural selection. The interactions of each species with its hosts produce a profusion of little “miracles of evolution”.

This charming and well-written volume gives a remarkably complete compendium of information on avian brood parasitism and comes up with answers to the major questions in parasitology and evolutionary biology that its study provokes. The book is filled with novel ideas and logical, incisive interpretations of empirical results. Every twist and turn of an argument is laid out clearly and precisely.

Davies begins by briefly introducing the six groups or lineages of obligate parasitic birds: the Old World and New World cuckoos, New World cowbirds, honeyguides, African finches and one South American duck. Each of these groups has attributes beyond their parasitic habits and adaptations that make them interesting. For example, the honeyguides in Africa have evolved a mutualism with honeybadgers and primates (including humans), in which they guide these mammals to honeybee nests. The mammal opens the nest, making the honey available to itself and the bee larva and waxy honeycomb available to the honeyguide (special bacteria in their gut allow the honeyguides to digest the beeswax).

Davies then delves deeper into the parasitic habits of these groups. He concentrates mostly on the Old World cuckoos and the cowbirds, the taxa for which we have the most detailed information. The book also contains a short and less satisfying section on conspecific brood parasitism (“Cheating on your own kind”). This topic really requires a book of its own.

Davies himself specializes in the common cuckoo (*Cuculus canorus*) of the Old World. Clearly, his best stories and speculations concern this species, and his treatment is comprehensive and enthusiastic. But what a remarkable species the common cuckoo is, with its host specificity, gentes (host ‘races’ based on egg coloration) and egg mimicry. There is the strange behaviour of its day-old nestling, which uses its back to push all of the host’s eggs or young out of the nest; and the ridiculous images of relatively massive cuckoo fledglings being fed by tiny warblers or robins.

Davies relates in detail the coevolutionary ‘arms race’ between the host (which should continuously evolve better defences against parasitism) and the cuckoo (which

should constantly counter with better ‘trickery’). For example, hosts should evolve the ability to recognize and reject cuckoo eggs that do not match their own. Many, but not all, hosts do this by ejecting the cuckoos’ eggs or deserting their nest.

In turn, rejection provides strong selection for cuckoos to evolve eggs that look like host eggs. In some cases, Davies points out, they may evolve eggs that are even more attractive to the host than the host’s own eggs! And, although the evidence for this is not definitive, it may be that the parasite places unique constraints on its host. The ‘mafia hypothesis’ involves cuckoo species whose nestlings do not typically eject host eggs. One of these, *Clamator glandarius*, will return to destroy the nests of hosts that reject its egg. This ‘offer you cannot refuse’ forces a host to accept the *Clamator* egg so as to avoid total loss of reproductive success (by accepting, at least some of its young may survive). Davies carefully presents the evidence for such adaptations and counteradaptations, and never oversimplifies or falls into the trap of describing ‘just-so’ stories.

Davies courageously and successfully describes the molecular methods that may elucidate how gentes have evolved in cuckoos and parasitic finches. He relates preliminary studies using DNA which show that gentes may evolve only along female lines (females are the heterogametic sex in birds), and may have evolved very recently in some groups of brood parasites. This is an important area of future research, from which we may be able to locate the particular sex-linked genes that control egg coloration, and quantify the sequence changes required to evolve gentes. The markers are also useful for determining the behaviour of individual female brood parasites.

The book is illustrated with lovely line drawings by David Quinn, but there are too



The stranger among us: the cuckoo's egg in this nest will present the brood with an unwelcome guest.

MIKE JONES/CORBIS

book reviews

few colour plates and their quality is patchy (the spotty egg photos are particularly disappointing given their importance in the book). The volume is written with both the professional scientist and interested amateur in mind, and would make an enjoyable read for both. Generally, this is a remarkable book, which describes brood parasitism and its evolution clearly and comprehensively, and sets the stage for the next round of exciting discoveries. ■

Robert C. Fleischer is at the National Zoological Park, Smithsonian Institution, Washington DC 20008, USA.

Time to deal with the legacy of secrecy

Undue Risk: Secret State Experiments on Humans

by Jonathan D. Moreno

W. H. Freeman: 2000. 320 pp. \$24.95, £14.95

Arthur C. Upton

This revealing account of government-sponsored experiments on human subjects benefits from Jonathan Moreno's experience as a senior member of President Bill Clinton's Advisory Committee on Human Radiation Experiments. Although Moreno makes no attempt to document such experiments comprehensively, the many examples he describes comprise a historically significant and illuminating series. They include research by Walter Reed and the Yellow Fever Commission on human volunteers at the beginning of last century, the Tuskegee syphilis study, experiments to find ways of protecting soldiers against mustard gas in the Second World War, and against chemical and biological weapons in the Gulf War, and other experiments to determine the effects of penicillin, nitrogen mustard, antimalarials, mescaline, plutonium and whole-body ionizing radiation.

The subjects in such research were patients, retarded children, volunteers, servicemen and prisoners, many of them recruited without informed consent, some unwittingly. For example, Moreno discusses the US government's failure to inform uranium miners, servicemen and members of the public about the risks of exposure to radiation when the United States was developing and testing atomic weapons. He also cites how NASA recruited Nazi experts to its space programme despite knowing about the atrocities perpetrated by the Nazis, and their horrific experiments on concentration-camp prisoners. Other egregious research discussed includes that conducted by Shiro Ishii on Chinese and Russian victims in Manchuria and by the Iraqi government on Kurds and other populations in Iraq.

In recounting details of these state-sponsored experiments, Moreno gives penetrating insights into the circumstances surrounding them. He also highlights the personalities involved, the issues that were at stake, and the personal, professional and public attitudes of the time. In the process, he chronicles the need for, and gradual evolution and acceptance of, an ethical framework for research on humans. Such a framework should embody the principles of voluntary participation and informed consent, as exemplified in the Nuremberg Code.

Moreno says at the end of the book that he believes experiments on humans will continue to be necessary, not only for the pre-market testing of new pharmaceuticals and other products, but also to protect populations from the threat of attack by atomic, chemical and biological weapons. He recommends that, in order to ensure that such experiments meet ethical standards of openness and informed consent, the international community should set up a "research ethics court".

In going behind the scenes to describe the struggles with the ethics of human experimentation among government officials concerned with national security, Moreno shows vividly how a policy of secrecy that seemed

justifiable to those directly involved proved ultimately to be corrosive to democracy and to seriously violate human rights. Looking at the lessons to be learnt from this, he argues persuasively that although research on humans for military and medical purposes will probably continue indefinitely, this should be only for defensive rather than offensive purposes, should meet ethical standards of informed consent, and should expose none of its participants to undue risks. Moreno's call for an ethical review process to ensure that investigators consistently meet such standards merits the serious attention of citizens and governments everywhere.

Undue Risk deals in a balanced, scholarly way with the issues involved, documenting the sources of information with appropriate references. Although the book was written for a general readership, it has a timely and important message, and should be relevant to all those concerned with military preparedness, defence and the ethics of human experimentation. ■

Arthur C. Upton is in the Department of Environmental and Community Medicine, University of Medicine and Dentistry of New Jersey—Robert Wood Johnson Medical School, 317 George Street, New Brunswick, New Jersey 08901, USA.

New in paperback

Evolving Brains

by John Allman

Scientific American Library, \$22.95

"It is quite remarkable in presenting a radically novel picture of the evolution of the brain..."

Evolving Brains is a masterpiece of synthetic presentation and evocative interpretation." Bob Martin, *Nature* 399, 426–427 (1997)

The Generation of Diversity: Clonal Selection Theory and the Rise of Molecular Immunology

by Alfred I. Tauber & Scott H. Podolsky

Harvard University Press, \$35.00, £21.95

"A dense and exhaustive scholarly treatise about the history of G.O.U.D. They explore the intellectual setting in which the Tonegawa experiment was performed by reviewing original literature and interviewing the major actors in the drama. This is not reading material for lay people, but everyone interested in the history of immunology and genetics should read this book." Fred S. Rosen, *Nature* 395, 235–236 (1998)

Burnet — A Life

by Christopher Sexton

Oxford University Press, £17.95

"I would have like to have seen a greater emphasis on the descriptions and implications of self-tolerance and clonal selection at the expense of detailing Burnet's more trivial encounters with

royalty. Nevertheless, the book brings alive the nature of the scientific quest and shows the excitement that a man of Burnet's genius can bring to the evolution of broad ideas, as well as how these ideas can spark off whole generations of scientists. Would that we could all have such an impact." I. M. Roitt, *Nature* 355, 779–780 (1992) — from the review of *The Seeds of Time: The Life of Sir Macfarlane Burnet*, the title under which the book was originally published in 1992.

The Nemesis Affair: A Story of the Death of Dinosaurs and the Ways of Science

by David M. Raup

W. W. Norton, £8.95

Avatars of the Word: From Papyrus to Cyberspace

by James J. O'Donnell

Harvard University Press, £9.50

The Bone Lady: Life As A Forensic Anthropologist

by Mary H. Manhein

Penguin Books, £11.95

Is the Temperature Rising? The Uncertain Science of Global Warming

by S. George Philander

Princeton University Press, \$16.95, £10.50

The tube worm turns

Science needs a new breed of Renaissance man and woman.

Roel Snieder

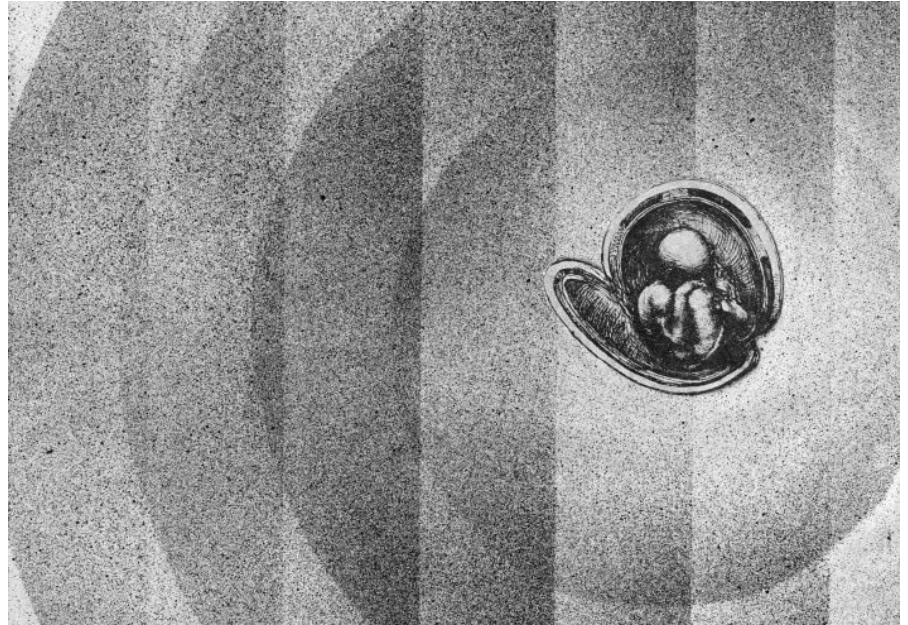
The past millennium has seen a dramatic change in the way scientists operate. During the Renaissance, scientists often had extremely broad interests. Leonardo da Vinci made great progress as a scientist and engineer, but he also was a painter, sculptor and architect. His interest in the anatomy of the human body was to grow into a *cosmografia del minor mondo* (a cosmography of the microcosmos). The phrase captures a connection between the macrocosmos and the biological world that is hard to find today.

In his work *Vetturale della Natura* (Conveyor of Nature), he made important contributions to fluid dynamics. As an engineer, he worked on the design of pumps, flying machines and other devices that formed an interface between his scientific interests and the needs of humankind. Because of his extreme breadth, Leonardo is the prime example of the *Homo universalis*. In the Renaissance, many scientists were also active in the field of religion; they often struggled with the conflict between their science and the religious views of that time. For some of them, such as Giordano Bruno, this struggle had a fatal ending.

Since the middle of the millennium, science has become increasingly specialized. These days, a scientist is a physicist, a biologist, a sociologist or other specialist who focuses primarily on her or his own field of interest. Even though there may be rare examples of scientists with an extremely broad area of expertise, *Homo universalis* has become virtually extinct.

This is not a positive development — specialization creates isolation. The scientific community can be compared to a population of tube worms. These animals live in colonies on the ocean floor at locations where hydrothermal vents release nutrients into the water. The tube worms are highly specialized; they survive only by extracting nutrients from the water at these vents. The worms cannot travel to other vents, which prohibits interbreeding among different colonies. Although they are well adapted to the harsh conditions of life on the ocean floor, they are completely reliant on the stability of their habitat. The specialization of tube worms has made them ill-equipped for new challenges.

But perhaps it is not quite fair to compare the scientific community to a collection of tube-worm colonies. There is now an increasing need to broaden the outlook of



Best of both worlds: Leonardo da Vinci meets modern remote-sensing techniques.

science and scientists once again. This development is spurred partly by the study of very complex systems. An example is the Earth's carbon-dioxide cycle, with its important implications for global climate. This cycle can be understood only by the combined efforts of chemists, biologists, physicists and Earth scientists. Because of the effect of industrial emissions on this cycle, this problem also has political, economic and social aspects. In addition to these complex problems, many fields that were traditionally separate are growing together. Biotechnology, granular media, the prediction of the world's population and optical computing are all areas that cannot thrive without the input of scientists from different disciplines.

In this sense, there seems to be a need to resurrect *Homo universalis*, but in a different form — not with the extreme breadth of a genius such as Leonardo da Vinci, but as a being who can work in an interdisciplinary team; who has the interests and the skills to 'look out of the box'; and who can com-

municate effectively with colleagues from other fields.

Many universities have recognized this trend and, because of present-day students' aversion to hard-core science, have developed broadly orientated programmes with titles such as 'science and society' or 'science and innovation'. But a scientist can make an innovative contribution to high-level interdisciplinary research only when she or he has mastered at least one discipline.

Interdisciplinary research should not be confused with 'no-disciplinary' research; in an interdisciplinary team, the contribution of each team member must be rooted in a professional knowledge of one research area. The failure of universities to acknowledge this obvious fact in the development of interdisciplinary teaching programmes leads to a watered-down version of a traditional scientific education. This is unfortunate because there is a real need for scientists with the skills and interests to interact with those from other fields.

Our challenge is to transform the scientific community from a collection of tube worms into an interacting web, resembling a diverse and rapidly evolving ecosystem where each member brings its own specialization to the party. This requires the resurrection of *Homo universalis* in a form appropriate for the twenty-first century. ■

Roel Snieder is in the Department of Geophysics and Center for Wave Phenomena, Colorado School of Mines, Golden, Colorado 80401, USA.

Homo universalis
needs to be
resurrected as a being
who can work in an
interdisciplinary team.

The flaw is human

The end of the Human Genome Reclamation Project

Jim Kling

Talbot Howard, Chair, North American Federation

From: Ronibus James, Senior Medical Advisor to the Chair of the North American Federation

Date: 25 March 2403

Dear Talbot,

You've asked me to give an official comment on requests for funding for research to re-introduce engineering of the human germline. I apologize for the delay in my response — the cystic fibrosis flared up again this week and the pulmonary treatments have put me behind in my work.

I have misgivings, Talbot. Not long ago, I was digging through the United States historical archive and some of the surviving political documents from that era, and I came across this exchange. I found it instructive, and perhaps you will, too. It is a transcript from a Senate budgetary committee meeting, dated 18 August 2238. Present are Alan Gillman, a Senator from Colorado, and the Chief Geneticist of the Human Genome Reclamation Project, Timothy Hopkins.

GILLMAN: Sir, your budget proposal contains a request for \$152 billion for the coming fiscal year, nearly one per cent of the federal budget. Surely you are aware that funding for agricultural and housing projects to cope with ongoing global warming leaves little room for superfluous expenditures. Why should we continue funding a program that seeks to resurrect genetic elements that we have spent eight generations expunging from the human gene pool?

HOPKINS: Sir, ecologists have been collecting and storing genetic data from endangered and extinct animal and plant species for years — surely you can see the value in a similar project dedicated to human beings? We can celebrate human diversity by establishing a database of human alleles, even harmful ones. These polymorphisms are the products of millions of years of evolution, Senator, and should be categorized and preserved with the same respect we give plant and animal species.

GILLMAN: That is a quaint notion, Mr Hopkins. But our forebears went to great expense and effort — in the face of no small societal resistance, I might add — to preserve us from genetic disease. Cancer, heart disease, cystic



fibrosis, sickle-cell anemia, all have been effectively eliminated. And yet you seek one per cent of the federal budget to flaunt humanity with its historical frailties, all for the sake of some genetic preserve.

HOPKINS: No, sir. I assure you we have no intent of producing human clones with these traits, only to store the alleles in a database — GILLMAN (*interrupting*): Frankly, I find such a notion offensive. We have achieved the greatest possible measure of health and contentment for the greatest possible number of human beings. Who are you to challenge the great work of your forebears?

HOPKINS: I don't challenge it, sir. We only wish to preserve the alleles that define what it is to be human.

GILLMAN: Surely by selecting dominant physical and intellectual traits, and removing abnormalities, we have produced the best of humanity.

HOPKINS: I disagree. There is no perfection. We are either well-adapted to our environment or we are not. We have indeed made ourselves an excellent match for the current environment, but the Earth is not a stagnant place. Things change.

GILLMAN: And what good could cancer genes do us, sir? Or genes that bring on heart disease, or predispositions to diabetes or stroke? Should things change, we will simply give our children the genes necessary to meet new challenges.

As the exchange suggests, the Senate budgetary committee chose not to continue funding of the Human Genome Reclamation Project, and it was disbanded in 2239.

As you are well aware, Hopkins' words were prophetic. Fourteen years later, the climatic and ecological changes brought on by

global warming unleashed an epidemic of Thai spotted fever that swept through the world's population, killing off nearly 98% in just three years. We have only recently recovered from the resulting breakdown in social order.

Most of the survivors of Thai fever were 'sports' — the term often applied to lower class citizenry, whose ancestral poverty had denied them the benefit of genetic manipulation of their bloodline. We now know that they carried recessive alleles that contributed to various genetic disorders. These had long been removed from engineered children, but in those less fortunate they continued to produce cases of Tay-Sachs disease, cerebral sphingolipidosis and many other disorders. In heterozygous individuals, however, they conferred resistance to Thai spotted fever.

Perhaps my opinion is already apparent. We should not seek again to engineer the human germline. Although cystic fibrosis and Tay-Sachs continue to be the leading causes of death in the North American Federation, we should fund only transient therapies. These conditions are our birthright — we renounce them at our peril.

Faithfully yours,

Ronibus James

Private addendum: on a happier note, Elizabeth gave birth this past Wednesday. Mother and daughter are doing beautifully. Elizabeth is heterozygous for CF, and we of course anxiously await the results of the tests.

Jim Kling is a science writer (<http://nasw.org/users/jkling>). He is writing a novel based in part on the premise of this story.

Crystals that breathe

Jonathan W. Steed

A non-porous crystal that can inhale and exhale sulphur dioxide gas without imploding offers a new way to make nanostructures. Microscopic gas sensors and switches could soon be created using this simple chemistry.

Everything seems to be 'nano' these days: nanotechnology, nanomachines, nanostructures... the list goes on. The nano-prefix refers to the nanometre unit of length, one billionth (10^{-9}) of a metre, and the implication is that the more 'nano' you can make technology, the better it will perform. This is already proving to be true for silicon chips, in which the smallest features of the next generation of processors will be around 10 nanometres wide. Building these 'nanostructures' requires laser-etching processes of increasing sophistication.

Laser etching, or lithography, is an example of what chemists call 'engineering down' a problem. In other words, as the size of the component decreases, the macroscopic technology needed to make it must become more refined. Even though it is enormously powerful, the 'engineering down' approach has several limitations, and it is nothing like the techniques used by living organisms to build their own components and molecular machines. The way in which natural components are made can be thought of as 'synthesizing up'. In this process — also known as 'self-assembly'¹ — sub-nanoscale molecular parts are brought together to produce nanometre-scale molecules.

A lot of research is dedicated to bridging the gap between these 'engineering down' and 'synthesizing up' approaches, and functional molecular machines are starting to be synthesized that can act as switches, sensors, levers, pulleys, gears, wires or rectifiers^{2,3}. The problem chemists have is not so much designing and building nanometre-scale molecules, but connecting these molecules (which are often dissolved in solution or available only as homogeneous powders) to the macroscopic world. After all, what use is a computer without keyboard, mouse and monitor screen?

The work of Albrecht and colleagues⁴, reported on page 970 of this issue, offers hope of a new 'synthesizing up' approach to nanotechnology. These researchers have carried out a rather simple piece of chemistry — the reversible binding of the gas sulphur dioxide (SO_2) by an organic molecule incorporating platinum ions (an organoplatinum compound). But they have done it in an ordered, solid crystal. It is the three-dimensional ordering of the process, bringing together nanotechnology and the macroscopic world, that is the remarkable feature of this work.

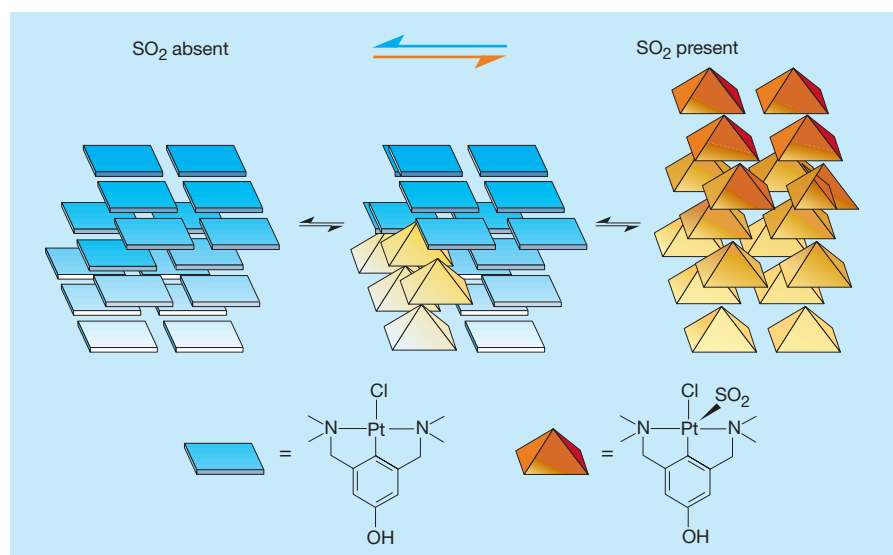


Figure 1 The organoplatinum crystal created by Albrecht and co-workers⁴. Normally, non-porous crystals cannot bind gas molecules without crumbling. But when the colourless organoplatinum crystal absorbs SO_2 , an orange colour spreads through the structure as the square platinum molecules transform into pyramids. With SO_2 present, the total volume of the crystal increases by 25%, but when the gas is 'exhaled' the crystal returns to its former size and colourless state. The crystalline structure remains intact throughout.

Not that the reversible binding of gases by solids is anything new. Porous materials such as zeolites and clays⁵ have been used for decades by the petrochemicals industry to bind all sorts of small molecules. Indeed, the binding ability of robust inorganic zeolites has spawned a whole host of designer porous structures that may be strong enough to withstand the formation of holes in their structure as molecular 'guests' enter and leave the 'host' framework⁶. Albrecht and colleagues⁴ now introduce a surprising new approach to this work — the idea that solid materials that can reversibly bind small molecular guests do not need to be porous.

Until now, it was assumed that any gross change in molecular structure and shape (such as that imposed by adding a molecule of SO_2 to an organoplatinum compound) would immediately result in fracture and crumbling of a crystalline solid, because of the extreme stresses and strains between the different-sized components. Furthermore, similar stresses and strains are known to accompany the removal of small molecular guests from ordered crystalline lattices, again resulting in crystals crumbling to pieces, or even imploding, as other molecules from the lattice rush to occupy the

empty space. The situation is summed up by the truism "nature abhors a vacuum".

Nature clearly does not abhor Albrecht and co-workers. When their colourless organoplatinum crystals are bathed in SO_2 for a minute or so, a beautiful orange colour develops. This spreads from the outside of the crystal inwards, as the square planes of the platinum molecules transform into the square pyramids of the SO_2 -crystal complex (Fig. 1). Some of the organoplatinum molecules have to wait patiently for their SO_2 partners while their neighbours expand and push the entire structure outwards. Even though the total volume of the crystal increases by about a quarter during the whole process, it remains perfectly three-dimensionally ordered. Even more remarkably, if the swollen crystal is exposed to air then the SO_2 is exhaled, with the crystal relaxing back down to its former colourless, SO_2 -free state (Fig. 1). The process can be repeated many times without loss of crystallinity. The organoplatinum crystals remain stable when exposed to both light and air, and seem robust enough to withstand the mechanical processing involved in device fabrication.

Albrecht and colleagues suggest that their

SO₂-sensitive material might make a crystalline optical switch because of the colour changes between the 'off' (SO₂ free) and 'on' (SO₂ present) states, or by virtue of its change in size. Such switches could be used to sense the presence or absence of SO₂. Moreover, the SO₂ switching could have wider applications in opto-electronic computational devices, if the crystals can be made to switch following interaction with light. They might also be used to modify laser frequencies. Access to a wide variety of laser frequencies is itself important in the lithographic manufacture of devices because the highest laser frequency (smallest wavelength) determines the size of the smallest features on a device or chip.

But why is this organoplatinum system so different? Why do these crystals 'breathe' when so many others crumble or simply do not react as solids? The answer is not yet clear but must reside in the interactions between one organoplatinum molecule and its neighbours. The crystalline framework is held together by a linear string of hydrogen bonds (weak links between adjacent molecules that involve a hydrogen atom on one of the molecules), which can tolerate a considerable degree of deformation. Moreover, the flat organoplatinum molecule may be able to tilt to accommodate the stress associated with the SO₂-induced expansion. Whatever the molecular reason for this crystal's remarkable properties, the most important aspect of this research is how it will affect the quest for

new sensors and nanoswitches. If organoplatinum complexes can 'breathe' SO₂, might some other simple compound work with chlorine, carbon dioxide, oxygen or even methane?

The success of this organoplatinum material suggests that other compounds that bind gases in solution may be capable of similar behaviour in the solid state. For example, would single crystals of haemoglobin be able to absorb and desorb oxygen? This essential biological molecule carries oxygen in the blood. Haemoglobin in solution reversibly binds oxygen in the same way that Albrecht and co-workers' organoplatinum crystal reversibly binds SO₂, despite having a very different structure. In the solid state, crystalline haemoglobin is highly hydrated and it is possible that gases could diffuse into it readily. Clearly the search for solid-state switches involved in binding and signalling is on. ■

Jonathan W. Steed is in the Department of Chemistry, King's College London, The Strand, London WC2R 2LS, UK.
e-mail: jon.steed@kcl.ac.uk

1. Philip, D. & Stoddart, J. F. *Angew. Chem. Int. Edn Engl.* **35**, 1154–1196 (1996).
2. Collier, C. P. et al. *Science* **285**, 391–394 (1999).
3. Steed, J. W. & Atwood, J. L. *Supramolecular Chemistry: An Introduction* (Wiley, Chichester, 2000).
4. Albrecht, M., Lutz, M., Spek, A. L. & van Koten, G. *Nature* **406**, 970–974 (2000).
5. Atwood, J. L., Davies, J. E. D., MacNicol, D. D. & Vögtle, F. (eds) *Comprehensive Supramolecular Chemistry* Vol. 7 (Pergamon, Oxford, 1996).
6. Yaghi, O. M. et al. *Acc. Chem. Res.* **31**, 474–484 (1998).

cause cancer when they are mutated in the cells of adults.

The *patched* gene encodes a large transmembrane protein that, in complex with the protein Smoothened, forms a receptor for the Hedgehog protein. In the absence of Hedgehog, Smoothened and Patched form an inactive complex in which Patched inhibits Smoothened. When Hedgehog binds, Smoothened is released from the clutches of Patched and sends a signal to the nucleus, leading to certain genes being switched on or off. Mutations in *patched* also switch on the Hedgehog pathway, resulting in molecular alterations that are related to cancer development in adult cells. For example, GLI, a protein further on in the pathway, is activated by mutations in *patched*. This promotes cell proliferation and has been implicated in brain tumours and other cancers. Nearly all basal-cell carcinomas have mutations in *patched*, and it is likely that activation of the Hedgehog pathway is necessary for skin cancer to develop³. So one approach to treating these cancers in adults might be to develop a compound that switches off the Hedgehog pathway. This is just what Taipale et al.¹ have done, with some help from an unusual source.

If Gorlin's syndrome can be thought of as the result of excessive activation of the Hedgehog pathway, then its molecular flip



Figure 1 The corn lily, *Veratrum californicum*. This lily contains the compound cyclopamine. When eaten by pregnant sheep, cyclopamine can produce abnormalities in their embryos. The same abnormalities can be produced by inactivation of the Hedgehog signalling pathway. By contrast, this pathway is activated in most human basal-cell skin cancers. As shown by Taipale et al.¹, cyclopamine might be useful in treating these cancers, by inhibiting the Hedgehog pathway.

Cancer

Sheep, lilies and human genetics

Allen E. Bale

One way of developing new treatments for cancer is to figure out the underlying molecular mechanisms of the disease, and then design drugs that target those mechanisms. Drugs that take advantage of the fundamental differences between cancer cells and normal cells — either killing cells that have tumour-specific defects or biochemically correcting the defects — are being developed or are in clinical trials. This has proved to be a slow and painstaking process. In the future, no doubt, systematic methods will be developed that can predictably translate knowledge of molecular mechanisms into useful drugs. But at the moment serendipity still has a part to play in finding both drugs and targets. On page 1005 of this issue, Taipale and colleagues¹ describe how they used the results of an epidemiological investigation of malformed sheep to identify a compound that might be effective in treating basal-cell skin carcinomas — the most common form of human cancer.

Basal-cell carcinomas typically occur in fair-skinned people beyond the age of 40. These tumours are rarely lethal because they grow slowly and do not migrate from the skin to internal organs. But they do invade surrounding tissues, and treatment — by surgery — can lead to disfigurement.

The molecular mechanisms that cause normal skin cells to transform into basal-cell carcinomas were deduced from studies of a rare hereditary disorder, Gorlin's syndrome, which is characterized by hundreds of these tumours². The gene mutated in this syndrome is a close relative of a gene involved in the development of the fruit fly *Drosophila melanogaster*. Called *patched*, this gene functions in the Hedgehog signalling pathway, which is involved in laying down the framework of a developing embryo. At first glance, the relationship between embryonic development and human cancer may not be obvious. But in fact many of the genes essential for normal development

side is holoprosencephaly — a disease associated with mutations in the human version of the fruitfly's *hedgehog* gene that inactivate it^{4–6}. The Hedgehog pathway is important in the development of the central nervous system and face, and holoprosencephaly stems from a failure of the embryonic brain and face to divide into symmetric halves. An extreme characteristic of this disorder is cyclopia, in which the two eyes are fused into one central structure and the remaining facial features and brain are rudimentary.

Like many other birth defects, cyclopia can be caused by either intrinsic genetic defects or the effects of environmental substances on genetically normal embryos. One such substance was discovered through epidemiological investigations of cyclopia in sheep herds of the western United States⁷. Sheep cyclopia was found to result when pregnant females ate a lily (*Veratrum californicum*; Fig. 1), and the relevant chemical from the lily was dubbed 'cyclopamine'. Although cyclopamine has no apparent effect on adult animals, it consistently leads to severe holoprosencephaly in developing embryos. As this condition can be caused by either mutations in human *hedgehog* or exposure to cyclopamine, it was suggested^{8,9} that cyclopamine acts by repressing the Hedgehog pathway. Taipale *et al.*¹ wondered whether cyclopamine might therefore be effective in treating basal-cell carcinomas.

For a compound to be effective against these cancers, it must switch off the Hedgehog pathway at a point downstream from the molecular defect — for example, downstream of mutant Patched protein. Taipale *et al.* now show that cyclopamine reverses activation of the pathway downstream of Patched and upstream of GLI. In fact, the evidence points to Smoothened as the site of the chemical's action.

The authors¹ also show that cyclopamine may be useful for treating patients with basal-cell carcinomas. These tumours grow slowly and are difficult to establish in culture. As a substitute, Taipale *et al.* used cells that were genetically engineered to lack a functional copy of *patched*. Using doses of cyclopamine that do not affect normal cells, they found that the genetically engineered cells stopped growing and that several of their malignant characteristics were reversed.

Together with the epidemiological finding that adult sheep do not suffer ill effects of cyclopamine, these results¹ are encouraging. But they do not directly address the question of whether switching off the Hedgehog pathway will cure basal-cell carcinomas. These tumours, like most other human cancers, have mutations in many genes, and it is probably the accumulation of mutations that results in malignancy. Perhaps the Hedgehog pathway must be activated to get tumour development started, but it is not certain that it needs to be kept activated to maintain

a fully malignant state. Indeed, in naturally occurring basal-cell carcinomas, shutting off this pathway might be like shutting the barn door after the horse has bolted. Treating basal-cell carcinomas with cyclopamine might be predicted to suppress the development of tumour cells and perhaps retard their growth, but not necessarily to kill them. There is some evidence that tumour cells that have been suppressed in this way might undergo programmed cell death or be eradicated by the body's natural defences. If not, then tumours treated with cyclopamine may eventually reappear when the drug is removed.

Fortunately, there is a good mouse model for Gorlin's syndrome that will help us to answer questions about the effects of

cyclopamine on tumours *in vivo*. If this drug then progresses to clinical trials, investigators would be wise to remember its source, and to avoid women of child-bearing age in their studies. ■

Allen E. Bale is in the Department of Genetics, Yale University School of Medicine, 333 Cedar Street, PO Box 208005, New Haven, Connecticut 06520-8005, USA.

e-mail: allen.bale@yale.edu

1. Taipale, J. *et al.* *Nature* **406**, 1005–1009 (2000).
2. Hahn, H. *et al.* *Cell* **85**, 841–851 (1996).
3. Sidransky, D. *Nature Genet.* **14**, 7–8 (1996).
4. Chiang, C. *et al.* *Nature* **383**, 407–413 (1996).
5. Belloni, E. *et al.* *Nature Genet.* **14**, 353–356 (1996).
6. Roessler, E. *et al.* *Nature Genet.* **14**, 357–360 (1996).
7. Keeler, R. F. & Binns, W. *Teratology* **1**, 5–10 (1968).
8. Cooper, M. K. *et al.* *Science* **280**, 1603–1607 (1998).
9. Incardona, J. P., Gaffield, W., Kapur, R. P. & Roelink, H. *Development* **125**, 3553–3562 (1998).

Artificial life

From robot dreams to reality

Rodney Brooks

Artificial life is a diverse field of research, but a common theme is teasing out the fundamental principles of life by building detailed working models. One of the most ambitious goals of artificial-life research is the construction of living systems out of non-living parts. Most of the artificial systems built remain strictly inside a computer, but on page 974 of this issue¹ Hod Lipson and Jordan Pollack take a first step towards bridging the gap between computer models and physical reality. They describe a system that evolves locomotive machines inside a computer and then automatically builds them, using rapid-prototyping technology, so that they can move around in the physical world.

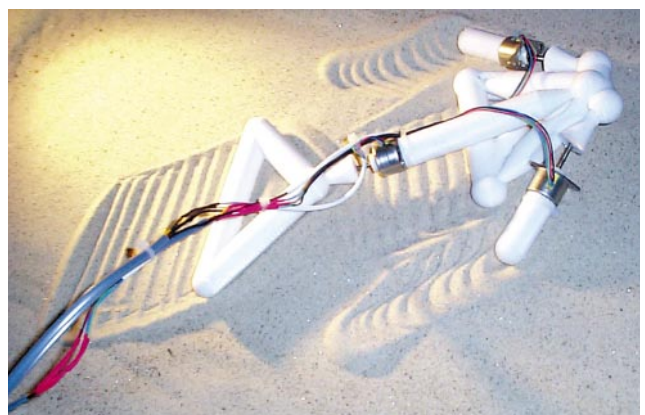
Over the past 20 years, computer algorithms inspired by genetics² (genetic algorithms) have become common tools in

solving optimization problems. Usually the optimization target is a mathematical procedure of known form, but with many undetermined parameters. A population of candidate procedures is generated by expressing different sets of these parameters as binary strings. The performance of each procedure is evaluated and those that are 'fitter' or do better according to some specific criteria are allowed to reproduce. Reproduction may be sexual, by taking the strings of two fit parent procedures and generating the binary string for the descendant by using cross-over and perhaps mutation, or it may be asexual, in which case just mutation is used. Over many generations, the population tends towards more successful procedures.

Over the past decade many people have experimented with evolving populations of 'artificial creatures' in simulated environ-

Figure 1 A robot 'evolved' by Lipson and Pollack¹ to produce horizontal motion. The body parts and control circuits are evolved inside a computer — thousands of robots over thousands of generations — and then rapid-prototyping technology is used to turn them into reality. Some of the winning designs are surprisingly symmetrical, which may be explained

by symmetrical machines finding it easier to move in straight lines. This particular machine uses antiphase synchronization to move — while the upper two limbs push the machine forwards, the central body is retracted, and vice versa. Movies of this robot and others are available at <http://www.nature.com> as Supplementary Information¹.



ments or virtual worlds. Here, rather than the binary string representing the parameters of some procedure, the string is an artificial genome, which encodes a control circuit (or nervous system) for a simulated robot. So, over time, better-performing robots slowly emerge. In some cases there is an implicit fitness evaluation as creatures fight it out for virtual resources necessary for survival. In other cases there is an explicit 'fitness function' applied to each generation of creatures, forcing evolution in a desired direction.

There are also experiments in transferring robot control systems evolved inside a computer onto physical versions of the simulated robots. But there is much debate and conflicting evidence over how well these transfer experiments work. In some cases³, the 10,000 or so fitness evaluations over tens to hundreds of generations for a population size of ~100 have all been done on physical robots — but such experimental tenacity is understandably rare.

A glaring omission from most of these experiments is that the 'body' of the robot is usually considered to be constant while just the nervous system evolves. An inspiring exception is the work of Karl Sims⁴. He evolved creatures that could swim and crawl in a virtual world that follows the laws of newtonian physics. His fitness function rewarded horizontal motion, and out popped creatures with locomotive ability, although some tweaking was required to get everything to work as planned.

An early version of the fitness function did not penalize vertical motion and was only applied to a few seconds of existence, whereupon really tall creatures evolved that were good at falling over and even tumbling. For a while, creatures evolved that moved along by beating their bodies with their limbs — they were taking advantage of a bug in the part of the simulated physics that encoded conservation of momentum. Sims' system coevolved nervous systems and body plans, but his creatures were all purely computational.

A couple of years ago Paolo Funes and Jordan Pollack⁵ tried to convert computer models into physical reality by evolving not creatures, just structures — the simulated structures were selected for their strength. They then built physical versions of the structures by hand from real Lego blocks and confirmed that the structures were much stronger than human-designed ones.

Lipson and Pollack now take this idea a step further. They evolve locomotive systems in computational space and use rapid-prototyping technology to automatically produce multi-linked structures that only need motors to be snapped on by hand. The successful designs evolved surprisingly different ways of generating motion, but there is reasonable correlation between the predicted locomotive abilities of the models and those measured from the physical robots (Fig. 1).

The evolution in Lipson and Pollack's experiments goes on within their computer — there is no fitness evaluation in the physical world. And the computational parts of the robots stay within that same computer even when they have been physically built. This means there can be no feedback from the physical world into the evolutionary process. At best, this system is like a virus that uses other more complex machines (which in this case are not life-forms themselves) to carry out reproduction. There is some way to go before self-reproducing robots can exist in the real world. But for artificial-life researchers there are aspects of the present study that satisfyingly resonate with the ways in which living systems develop.

First, these particular robots cannot be built by conventional manufacturing techniques. The rapid-prototyping technology solidifies polymers in place, so that ball and socket joints are constructed with the ball already inside the socket. The parts are never separate, and if they were they could not be assembled without damaging them. This is not far removed from the way that biological systems grow. Second, the evolutionary strategy used in these experiments starts with a blank or 'null' genome and randomly mutates it into one that generates a working

machine — so there is no in-built bias from seed machines in the population. One could say these machines have evolved 'naturally', without human intervention.

People have long thought about building technology from non-biological materials that can reproduce themselves — the US space agency NASA convened a panel in the 1960s to investigate the possibility of seeding the Moon with a small self-reproducing factory. Although we are still a long way from that goal, Lipson and Pollack have at last demonstrated a computational system that designs functional machines and builds them with almost no human intervention. The resulting machines cannot match the complexity of the rapid-prototyping machine designed by human engineers that is required to do the actual fabrication. Nevertheless, this is a long awaited and necessary step towards the ultimate dream of self-evolving machines. ■

Rodney Brooks is in the MIT Artificial Intelligence Lab, 545 Technology Square, Cambridge, Massachusetts 02139, USA.

e-mail: brooks@ai.mit.edu

1. Lipson, H. & Pollack, J. B. *Nature* **406**, 974–978 (2000).
2. Holland, J. *Adaptation in Natural and Artificial Systems* (Univ. Michigan Press, 1975).
3. Cliff, D. et al. *Adapt. Behav.* **2**, 73–110 (1993).
4. Sims, K. *Artif. Life* **1**, 353–372 (1994).
5. Funes, P. & Pollack, J. *Artif. Life* **4**, 337–357 (1998).

Bacterial genomics

Pump up the versatility

E. Peter Greenberg

On page 959 of this issue, Stover and colleagues¹ publish the genome sequence of the resilient bacterium *Pseudomonas aeruginosa*. This microbe is feared by every patient with the genetic disorder cystic fibrosis because it colonizes the lungs of most people with this disease. Once established, it slowly but surely causes more and more damage to the lungs, eventually leading to the patient's death. This opportunistic pathogen also infects people whose immune system is compromised in any way. It rarely infects uncompromised hosts, and occurs naturally almost everywhere — in lakes, streams, soil and even our drinking-water supply. *P. aeruginosa* (Fig. 1, overleaf) shows great nutritional versatility; like a goat, it can live on almost anything.

Traditional approaches have not provided us with the tools needed to fight this bacterium effectively. For reasons that are not clear, *P. aeruginosa* infections resist treatment with antibiotics. Hence the hope that the *P. aeruginosa* genome sequence¹ will reveal new ways of tackling this organism.

The completed sequence, described by Stover and colleagues, is the result of a unique collaboration between a group of

academic investigators, a pharmaceutical company and a charitable foundation, the US Cystic Fibrosis Foundation. The project, which began three years ago, was financed jointly by the Cystic Fibrosis Foundation and the company PathoGenesis, and has proven to be a model of scientific integrity and interaction.

Academics at the University of Washington in Seattle sequenced and analysed the genome, using the 'shotgun' sequencing and compiling strategy developed by Craig Venter and colleagues². Periodically, the available sequence data were posted on the Internet³. Scientists at PathoGenesis — with the help of 61 *P. aeruginosa* experts from around the world — then 'annotated' the genome. What this means is that they picked out sequences in the genome that were likely to be genes and used bioinformatics to compare these predicted genes with known genes from *P. aeruginosa* and other species. This allowed them to work out what the functions of the *P. aeruginosa* genes might be.

The collaborators hoped to gain insights into the basic biology of *P. aeruginosa*, to find targets for drug development, and to spur interest among scientists in studying this

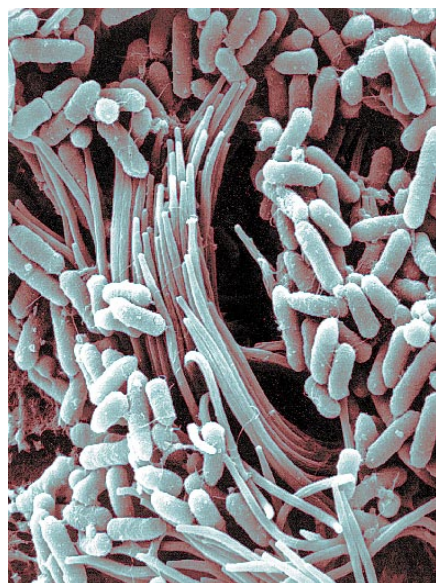


Figure 1 The rod-shaped *Pseudomonas aeruginosa* on cultured epithelial cells from the human respiratory tract. Each *P. aeruginosa* cell measures about 0.5 by 2–3 μm . The bacterium shows extraordinary nutritional versatility. This might require the large number of genes in its genome (the sequence of which is now published¹) that are predicted to be involved in regulating gene expression.

resilient beast. On all counts, the project has met its goals.

Pseudomonas aeruginosa has a large genome, about one-third larger than that of *Escherichia coli*. In fact it has about as many genes as the budding yeast *Saccharomyces cerevisiae*, and about half as many as the fruitfly. The large size of the *P. aeruginosa* genome allows for a wide diversity of genes. Although this is no great surprise, the more detailed annotation of the genome has been quite revealing. Two features that jump out can be summarized as regulation, regulation and more regulation, and pumps, pumps and more pumps.

Stover *et al.* discovered a large number of genes — between 8% and 10% of the predicted total of 5,570 genes — that encode proteins with sequences similar to those of known regulators of gene expression. The authors note a correlation between genome size and the percentage of predicted regulators in sequenced bacterial genomes: the larger the genome, the higher the percentage of regulatory genes. For example, *Helicobacter pylori* — a highly specialized bacterial pathogen that can colonize the acidic environment of the human stomach — has 1,709 predicted genes. Fewer than 20 of these (about 1%) are expected to be regulatory. Smaller genomes have even lower percentages of regulatory genes. The incredible potential for *P. aeruginosa* to turn genes on and off according to the conditions in which it finds itself is consistent with its nutritional versatility.

The *P. aeruginosa* genome also encodes

a number of pumps, which may be the key to the bacterium's ability to withstand antibiotics in its human host. It has been suggested that Gram-negative bacteria, a group to which *P. aeruginosa* belongs, can resist the lethal effects of many antibiotics by pumping them out of their cells faster than the chemicals can accumulate. Members of the RND protein family of 'multi-drug efflux' pumps can render bacteria impervious to antibiotics⁴. It is not clear how or why these pumps evolved; perhaps they were first needed to eliminate naturally occurring environmental toxins. We already knew of four RND pumps in *P. aeruginosa* (*E. coli* has the same number, whereas *Mycobacterium tuberculosis*, the persistent pathogen that causes tuberculosis, has none). The sequencing project has now revealed six more.

When and where are these pumps active? One idea is that they might be expressed when *P. aeruginosa* exists as a biofilm. These are stationary communities of bacteria that are enclosed by a self-produced extracellular matrix. When bacteria exist as biofilms, they can resist antimicrobial agents. A hypothesis to explain the persistence of *P. aeruginosa* in the lungs of a cystic fibrosis patient is that it may establish a biofilm in this environment⁵.

Mathematics

The Lorenz attractor exists

Ian Stewart

For nearly 40 years, one of the classic icons of modern nonlinear dynamics has been the Lorenz attractor. With its intriguing double-lobed shape and chaotic dynamics, the Lorenz attractor has symbolized order within chaos (Fig. 1). The only problem is: does it exist? Mathematicians have lacked a rigorous proof that exact solutions of the Lorenz equations will resemble the shape generated on a computer by numerical approximations, and they also could not prove that its dynamics are genuinely chaotic. Perhaps the calculations showed something that merely looked like chaos — a numerical illusion. The smart money has always been on chaos in the Lorenz system being real, but the rigorous techniques of dynamical mathematics were unable to prove it. Until last year, that is, when Warwick Tucker of the University of Uppsala completed a PhD thesis showing that Lorenz's equations do indeed define a robust chaotic attractor. The proof has since been published (W. Tucker, *C. R. Acad. Sci.* **328**, 1197–1202; 1999), and an excellent summary has been provided by Marcelo Viana (*Math. Intell.* **22**, 6–19; 2000).

Tucker's work is hugely significant, not

The abundance of regulatory genes and questions about the regulation of antibiotic pumps in *P. aeruginosa* call for investigators to look into gene expression in this bacterium. To this end, the US Cystic Fibrosis Foundation is again taking a unique approach, by building on its initial investment in *P. aeruginosa* genomics and underwriting the cost of constructing a *P. aeruginosa* gene-expression array. The charity plans to make the array available to interested researchers at a cost consistent with the budgets of academic scientists. It has also created a Cystic Fibrosis Bioinformatics Center, to support scientists using the arrays. One hopes that the genomic information reported by Stover *et al.*¹, and the information to come from the gene-expression analysis, will allow us finally to tame this testy pathogen.

E. Peter Greenberg is in the Department of Microbiology, College of Medicine, University of Iowa, Iowa City, Iowa 52242-1109, USA.

e-mail: everett-greenberg@uiowa.edu

1. Stover, C. K. *et al.* *Nature* **406**, 959–964 (2000).
2. Fleischmann, R. D. *et al.* *Science* **269**, 496–512 (1995).
3. The *Pseudomonas* Genome Project.
<http://www.pseudomonas.com/>
4. Nikaido, H. *Clin. Infect. Dis.* **27**, S32–S41 (1998).
5. Costerton, W. C., Stewart, P. & Greenberg, E. P. *Science* **284**, 1318–1322 (1999).

just because it provides the Lorenz attractor with a solid foundation, but because his techniques will be widely applicable. At last, the embarrassing gap between what we think we know about a nonlinear dynamical system from numerical simulations, and what we actually know in full logical rigour, is starting to close. At the moment the methods are limited to dynamics in three dimensions, but after Tucker's breakthrough that may well change. Dimensions greater than three are of considerable interest because the dimension of a dynamical system is not the dimension of ordinary space, but the number of variables in the equations. For example, the motion of a three-body system consisting of the Earth, Mars and a space probe requires six variables for each body — three of position and three of velocity — and so is an 18-dimensional dynamical system.

Why bother with rigorous proofs? Surely any practical implications of the equations are already embodied in the numerical results — do we need to be obsessed with logical rigour? Yes, we do. There are several reasons for taking numerical solutions of nonlinear differential equations with a pinch of salt. Numerical methods are approxima-

tions, and chaotic systems are highly sensitive to approximations; it is well known that numerics can sometimes give seriously misleading results. But the deepest reason is that until we can prove what our computers seem to be telling us, then we are ignoring a huge gap in our mathematical technique. Often such a gap is a clue that important conceptual ideas are lurking nearby, as is the case here.

The Lorenz attractor dates from 1963, when the meteorologist Edward Lorenz published an analysis of a simple system of three differential equations that he had extracted from a model of atmospheric convection. He pointed out that they possess some surprising features. In particular, the equations are 'sensitive to initial conditions', meaning that tiny differences at the start become amplified exponentially as time passes. This type of unpredictability is a characteristic feature of chaos. Conversely, there is also 'order' in the system: numerical solutions of the equations, plotted in three dimensions, consist of curves winding round and round a curious two-sheeted surface, later named the Lorenz attractor.

The geometry of the attractor is closely related to the 'flow' of the equations — the curves corresponding to solutions of the differential equations. There is an unstable equilibrium, a saddle point, at the origin. The curves repeatedly pass this point, and are pushed away to the left or right, only to circle round to pass back by the saddle. As they loop back, adjacent curves are pulled apart — this is how the unpredictability is created — and can end up on either side of the saddle. The result is an apparently random sequence of loops to the left and right.

The central technical issue in proving that the Lorenz system is a chaotic attractor is translating these statements into suitably precise mathematics. Tucker's proof combines two main ideas. One is a conceptual characterization of chaotic attractors in terms of 'singular hyperbolicity', introduced by C. Morales, M. J. Pacifico and E. Pujals (*C. R. Acad. Sci.* **326**, 81–86; 1998). Previous work on dynamical systems had concentrated on 'hyperbolic' systems, where the flow of the equations can always be split into a set of contracting directions and a complementary set of expanding ones. But the Lorenz system is not hyperbolic. Singular hyperbolicity replaces the idea of expanding directions in the flow by the condition that part of the flow should expand in volume. If some sides of a box expand and others contract, the volume may nonetheless expand if the amount of expansion beats the amount of contraction. So singular hyperbolicity is less restrictive than hyperbolicity.

Tucker's other important idea is using computer calculations in a rigorous way to establish certain features of the geometry of the solutions of differential equations — normal numerics plus precise error esti-

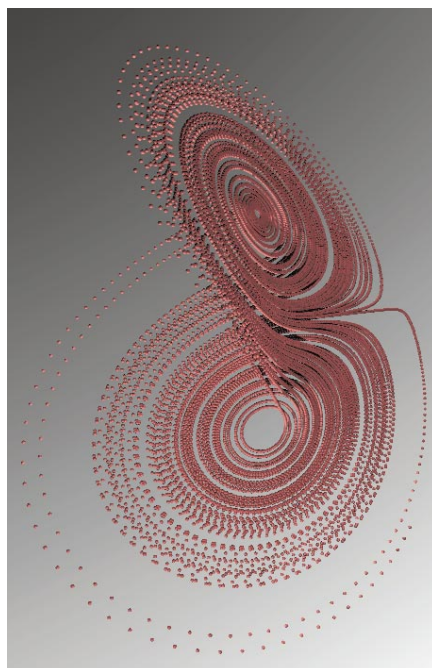


Figure 1 The Lorenz attractor was first derived from a simple model of convection in the Earth's atmosphere. Previously, the Lorenz attractor could be generated only by numerical approximations on a computer, as shown here. Now we have a rigorous proof that confirms its existence.

mates. Tucker obtains a rigorous approximation to the way curves loop back towards the origin by following the time evolution of lots of tiny boxes (up to 10,000). But numerical solutions of differential equations encounter problems near saddle points, because the flow slows down exponentially fast. To get round this difficulty, Tucker uses the well-

established technique of a 'normal form'. Near an equilibrium, any differential equation can be transformed into an equation that can be solved by an exact formula, to a high degree of approximation. This formula can be used instead of the numerical procedure whenever the flow gets too close to the equilibrium.

The task then reduces to finding a set of boxes such that curves starting within one box eventually return to a box (usually a different one), which in effect says that the collection of boxes forms an attractor. Some further technical conditions, related to singular hyperbolicity, are required to establish that the flow is chaotic. The search for suitable boxes begins with an informed guess based on raw numerics; any boxes that cause trouble by growing too fast are subdivided until they cease to cause trouble. When the subdivision process of all boxes stops, the proof is complete. Finding a suitable collection of boxes took about 30 hours on a fast computer.

Technicalities aside, the main idea here is that, with care, naive numerics can be used with precise error estimates to establish significant features of the flow of a nonlinear differential equation. When these features are combined with appropriate conceptual insights, the existence of chaotic attractors becomes irrefutable. So, thanks to Tucker, dynamical systems theorists can at last stop worrying about whether their most potent icon might suddenly fall apart. And Lorenz's original insight, that the strange behaviour of his equations was not a numerical artefact, can no longer be disputed. ■

Ian Stewart is at the Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK. e-mail: ins@maths.warwick.ac.uk

Malaria

Channelling nutrients

Kiaran Kirk

Malaria, an infectious disease responsible for an estimated 300 million to 500 million clinical cases and 1.5 million to 2.7 million human deaths each year¹, is caused by a single-celled parasite that invades the red blood cells of its host. In the 48 hours after it invades a red blood cell, the parasite grows to many times its original size, and then divides to produce 20–30 new parasites. To fuel this high rate of growth and multiplication, the malaria parasite needs nutrients from outside the infected cell. However, a normal red blood cell is unable to take up at least some nutrients fast enough to satisfy the parasite's voracious appetite. On page 1001 of this issue², Desai and colleagues provide new insights into how this problem is solved. They describe a versatile channel

that is found in the outer membrane of infected (but not uninfected) red blood cells, and which helps to supply nutrients to the hungry parasite.

It has long been known that following the invasion of a human red blood cell by *Plasmodium falciparum* — the most virulent of the four malaria parasites that are infectious to humans — the membrane of the infected cell undergoes a dramatic increase in its permeability to a range of small solutes, both charged and uncharged^{3–6}. The pathways responsible for the increased permeability have previously been shown to have a strong preference for negatively charged ions (anions) over positively charged ions (cations), and to be blocked by drugs known to inhibit anion-selective channels⁶

(pore-forming proteins that provide anion-selective pathways across cell membranes). This led to the proposal⁶ that the increased permeability is the result of the parasite activating or inserting an anion-selective channel in the membrane of the red blood cell.

Ion channels are best studied in detail using electrophysiological methods that allow the electrical current associated with the flow of ions through the open channel to be measured directly. One such method, the patch-clamp technique, involves attaching a glass pipette to a region on the surface of a cell and measuring the current flowing either across the entire cell membrane or through individual channels in the small patch of membrane enclosed within the glass pipette. Desai *et al.*² have used both approaches to obtain direct electrophysiological evidence for an unusual, anion-selective channel that is present in the membrane of red blood cells infected with *P. falciparum* but not seen in normal red blood cells. The fragility and small size of the parasitized red blood cell make it a difficult target for this type of experiment, and the work described by Desai *et al.* is a significant technical achievement.

The results show that the parasite-induced channel has a strong preference for anions over cations, and that it allows both inorganic and organic anions to move across the cell membrane. It is blocked by a range of compounds previously shown to inhibit the pathways responsible for the increased permeability of infected red blood cells. Moreover, estimates of the relative permeability of the channel to different solutes agree remarkably well with previous estimates — made using quite different methods — of the relative rates of uptake of these solutes into infected cells.

The permeability of the channel to organic solutes allows it to serve as a route for both the uptake of nutrients into the cell and the removal of potentially hazardous metabolic wastes. Anionic nutrients such as the amino acid glutamate² and the water-soluble vitamin pantothenate⁷ flow readily through the channel. Both are required by the parasite⁸, but neither passes easily across the membrane of a normal red blood cell, so the parasite relies on the new channel to bring them into the cell. The parasite-induced channel is also used for the uptake of neutral compounds, although at least some of these, such as D-glucose, also enter the infected cell rapidly through pathways that are native to the cell membrane. Lactate, a by-product of glucose metabolism, is produced in huge quantities by the parasite, and the channel provides a route for it to flow out of the cell^{2,9} down its concentration gradient.

The channel has a high permeability to chloride ions, which ensures that it does not perturb the normal voltage across the red-blood-cell membrane. Nor does it disturb the normal distribution of chloride ions —

which are at equilibrium — across this membrane. However, it has a very low permeability to sodium ions, and this is important. Sodium is the major cation in the blood plasma. If the new channel were to allow sodium to pass freely across the red-blood-cell membrane, sodium and chloride ions, together with water, would rush into the cell, causing it to swell and burst before the parasite had time to complete its growth and multiplication.

The work by Desai *et al.*² gives new insights into the properties of an unusual channel, which provides the intracellular malaria parasite with an effective and parsimonious way of moving important solutes into and out of its host cell without threatening the cell's stability. We do not yet know the molecular identity of the channel, nor whether it originates from the parasite or from the red blood cell itself. But we do know enough to recognize that the channel is an attractive target for new antimalarial drugs. Pharmaceutical companies have been showing increasing interest in anion-selective

channels as drug targets for the treatment of a range of diseases. Screening chemical libraries for molecules that block this channel may well lead to much-needed new drugs with which to combat the increasingly drug-resistant malaria parasite.

Kieran Kirk is in the Division of Biochemistry and Molecular Biology, Faculty of Science, Australian National University, Canberra 0200, Australia.
e-mail: kieran.kirk@anu.edu.au

1. Trigg, P. I. & Kondrachine, A. V. in *Malaria: Parasite Biology, Pathogenesis, and Protection* (ed. Sherman, I. W.) 11–22 (Am. Soc. Microbiol., Washington DC, 1998).
2. Desai, S. A., Bezrukov, S. M. & Zimmerberg, J. *Nature* **406**, 1001–1005 (2000).
3. Elford, B. C., Haynes, J. D., Chulay, J. D. & Wilson, R. J. *Mol. Biochem. Parasitol.* **16**, 43–60 (1985).
4. Ginsburg, H., Kutner, S., Krugliak, M. & Cabantchik, Z. I. *Mol. Biochem. Parasitol.* **14**, 313–322 (1985).
5. Gero, A. M., Buglédich, E. M., Paterson, A. R. & Jamieson, G. P. *Mol. Biochem. Parasitol.* **27**, 159–170 (1988).
6. Kirk, K., Horner, H. A., Elford, B. C., Ellory, J. C. & Newbold, C. I. *J. Biol. Chem.* **269**, 3339–3347 (1994).
7. Saliba, K. J., Horner, H. A. & Kirk, K. *J. Biol. Chem.* **273**, 10190–10195 (1998).
8. Divo, A. A., Geary, T. G., Davis, N. L. & Jensen, J. B. *J. Protozool.* **32**, 59–64 (1985).
9. Kanaani, J. & Ginsburg, H. *J. Cell. Physiol.* **149**, 469–476 (1991).

Geology

Probing the memory of mud

Alan Kemp

Mudrocks are some of the most common yet least glamorous geological formations. They erode easily so they are often eclipsed by more robust and photogenic outcrops of granite, sandstone or limestone. Yet mudrocks are an important archive of the Earth's history, and contain the detritus of the physical and biogeochemical cycles that shape and regulate global systems. A major constituent of mudrocks is quartz (or silica, SiO₂), one of the most abundant rock-forming minerals and the most resistant to weathering. Until recently, it had been assumed that the quartz found in these ancient sediments came from the erosion and fragmentation of quartz crystals from pre-existing rocks. On page 981 of this issue, Schieber and others¹ challenge this view and propose a radically different interpretation: that much of the quartz in mudrocks may come from the siliceous skeletons of marine plankton (Fig. 1, overleaf).

The implications of this study are wide ranging, from our understanding of the evolution of marine plankton and links between the biogeochemical cycles of silica and carbon, to the interpretation of fine-grained mudrocks and their use in reconstructing past environments. Because they are made of silt- and clay-sized particles (less than 62 and 4 micrometres, respectively), the analysis of mudrocks is beyond the resolution of optical microscopy and

requires more time-consuming and expensive techniques.

The work of Schieber *et al.*¹ is based on the innovative use of three complementary methods. Under a scanning electron microscope, the textures of individual particles are resolved using back-scattered electron imagery and cathodoluminescence. Non-luminescent quartz silts are characteristic of formation *in situ* (by dissolution and re-precipitation of SiO₂ within the sediment) rather than by erosion from pre-existing rocks. Schieber *et al.* identified most of the quartz silt in their mudrock samples as non-luminescent. To confirm this interpretation they used an ion microprobe to measure oxygen isotope ratios from individual quartz grains.

Until now, the proportion and size of quartz grains in mudrocks have been used to reconstruct the historical layout, or palaeogeography, of ancient seaways. This is because the proportion of land-derived detritus generally decreases with distance from shore and patterns of grain sizes relate to the speed and direction of ancient currents. These assumptions are no longer tenable for the Devonian (380-million-year-old) shale studied by Schieber *et al.* and must now be in doubt for other ancient mudrocks pending further ion-probe investigations of this type.

The findings also call into question



100 YEARS AGO

As attempts are being made to found a domestic science, and to introduce exactitude into the operations of the kitchen, a note in the Monthly Weather Review recording the actual experience of a housekeeper at Albuquerque, New Mexico, is of interest. It appears that cooking recipes and practices which are trustworthy not far from sea-level are worthless at Albuquerque, the altitude of which is 4933 feet. Water boils there at 202 °F, instead of 212 °F; hence articles of food, the cooking of which depends upon heat applied through the medium of water, require a longer time for cooking than is given in the cookery books... But the worst difficulty is with cake-making. Ordinary recipes as to number of eggs and amount of baking powder break down altogether, and housekeepers have to modify them if they wish their operations to be successful. As the barometric pressure determines to what extent the disengaged carbon dioxide shall expand and aerate the dough, this may explain the different action of baking soda and egg batter. In any case, the observation is interesting, and chemists may find it worthy of their attention.

From *Nature* 30 August 1900.

50 YEARS AGO

Maui-of-a-Thousand-Tricks, his Oceanic and European Biographies. By Katharine Luomala. From time immemorial adventure stories have taken hold of the imagination, and this is partly why the cult of Maui-of-a-thousand-tricks is so widespread in the Pacific. The origin of this hero is not unparalleled in other mythologies: born of human parents, he was brought up by the gods and learnt their magic, so that when he returned to earth he was able to use this knowledge in part to aid humanity, in part for his own diversion and pranks... The cult of Maui, though predominantly Polynesian, is found also in Micronesia and Melanesia, principally where there has been contact with Polynesia. In some islands there is a long story cycle, in others merely a footprint... there can be no 'true' account of this hero, but the fact remains that he won the hearts of the islanders over a vast region and maintained that hold for centuries... Katharine Luomala is to be congratulated on her perseverance in collecting this vast amount of material... the fact that the bibliography runs into more than three hundred items speaks for itself.

From *Nature* 2 September 1950.

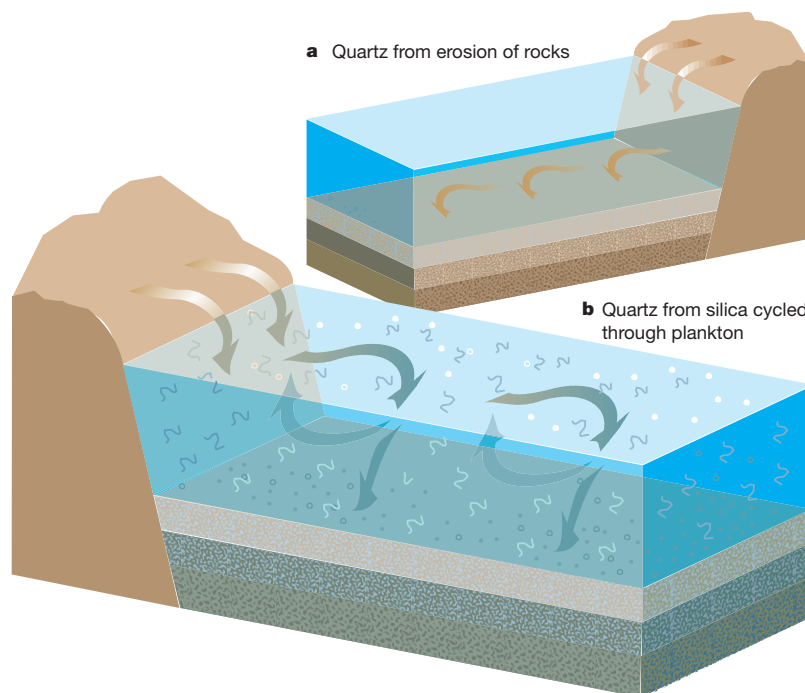


Figure 1 What is the origin of the quartz found in ancient mudrocks? Previously, most of the quartz was thought to be eroded from older rocks (a), but it could also come from the dissolved silica of plankton skeletons (b). A study by Schieber *et al.*¹ now favours the latter process as the main source of quartz in some mudrocks, although ultimately the silica in the plankton would have come from weathered rocks.

reconstructions of past climate in which large amounts of sedimentary quartz silt (thought to be carried by the wind) have been used to infer increased wind strength. Of course, to be present in the mudrocks at all, some SiO₂ must have been transported out to sea, but probably in its dissolved form, silicic acid, which is used by plankton to make skeletons, rather than as quartz grains. This in turn suggests that the main mechanism by which dissolved silicon is supplied to the oceans — the chemical weathering of rocks and subsequent transport by rivers² — was probably more intense than previously thought.

What, then, is the source of the quartz silt in these mudrocks? One might expect the sediment to contain remnants of ancient organisms, such as diatom algae that use SiO₂ for their hard shells. Unfortunately for those seeking to trace the fossil record of such organisms, their hard parts are made of opal — the hydrated form of silica (SiO₂·0.4H₂O). Opal is structurally amorphous and unstable, and transforms rapidly on burial to crystalline quartz, destroying any delicate skeletal structures. But Schieber *et al.* did find evidence for radiolarians (unicellular protozoan zooplankton) in their samples. Some of their opaline skeletal structures were preserved in nodules that acted as sealed microenvironments and so stopped them from dissolving completely, yet it seems unlikely that radiolarians alone could have contributed all the quartz formed *in situ*.

The oldest known fossil diatom is around 190 million years old³, but there is molecular evidence that suggests that diatoms first appeared 400 million years ago⁴. Schieber *et al.* speculate that the discovery of abundant quartz silt formed *in situ* may explain the 'missing' diatom fossil record, in so far as the silica from the vanished diatoms may have been transformed into quartz silt. But the reliability of the 'molecular clock' is no longer certain⁵ and a more conservative molecular reconstruction gives a much younger date (266 million years ago) for the origin of diatoms⁶, which is much closer to that of the existing fossil record.

If the quartz silt is not derived from diatoms, the next best candidate may be a diatom precursor, originally a naked cell that acquired a coating of siliceous scales⁷. Siliceous scales only a few micrometres in size, resembling those of modern chrysophyte algae, have also been found in much older rock strata (from 550 million years ago)⁸. Whatever the origins of the *in situ* quartz, the answer lies in further investigations of ancient mudrocks.

Last month, an article in these pages⁹ drew attention to the close links between the biogeochemical cycles of silica and carbon over recent millennia through the key role played by siliceous plankton in the biological carbon pump (the flux of carbon to the sea floor). Significantly, the sediments studied by Schieber *et al.* are also rich in organic carbon. The actions of biogeochemical cycles

can be paraphrased by the words of Dylan Thomas¹⁰: "The force that through the green fuse drives the flower". Here, the force is the solar energy that drives photosynthesis. The flower, in this case the organic carbon of the phytoplankton, sinks from the surface waters drawing down carbon dioxide from the atmosphere and, if buried in sufficient quantities in the sediment, creating oil-rich rocks. For the past 100 and possibly 200 million years, diatom algae have been the primary 'green fuse' in this process. If we are to adequately understand the workings of biogeochemical cycles in earlier periods, then identifying the 'green fuse' of the past is essential. ■

Developmental biology

Giving limbs a hand

Martin J. Cohn

Vertebrate limbs are complicated structures. The genetic programme that directs their development is also proving to be complex, and trying to understand how it works is keeping developmental biologists busy. Three papers in *Development*^{1–3} now bring the molecular basis of limb development closer to hand, or rather to *dHAND*, a gene with a key role very early on in the process.

As three-dimensional structures, limbs develop along three axes. In the arm these are: shoulder to fingertips, thumb to small finger, and palm to back of the hand. All limbs start out as limb buds in an embryo. Each bud is an undifferentiated swelling made up of only two kinds of cells — a core of mesenchymal cells, and a surrounding jacket of epithelial cells. The challenge for the embryo is to transform the limb bud into the elaborate system of different tissues found in the mature limb. This patterning process is orchestrated by specialized groups of cells that act as signalling regions and tell other cells where they are and what to do. For example, patterning along the anterior to posterior (thumb to small finger) axis is controlled by mesenchymal cells at the posterior edge of the limb in an area known as the polarizing region. Here, cells secrete the signalling molecule Sonic hedgehog (*Shh*), a protein that dictates the identities of skeletal elements and is needed for the limb to develop (Fig. 1). But what is signalling to the signaller? In other words, how is the polarizing region established?

The *dHAND* gene (also called *HAND2*) encodes a transcription factor, the dHAND protein, best known for its involvement in heart and face development^{4,5}. It now turns out that this protein is also required for the development of paired appendages (fins, as well as limbs; Box 1, overleaf), where it either

Alan Kemp is in the School of Ocean and Earth Science, University of Southampton, Southampton Oceanography Centre, European Way, Southampton SO14 3ZH, UK.
e-mail: aesk@soc.soton.ac.uk

1. Schieber, J., Krinsley, D. & Riciputi, L. *Nature* **406**, 981–985 (2000).
2. Tréguer, P. et al. *Science* **268**, 375–379 (1995).
3. Rothpletz, A. Z. *Deutsche Geol. Ges.* **48**, 184–194 (1896).
4. Philippe, H. et al. *Evol. Biol.* **7**, 247–265 (1994).
5. Ayala, F. J. *Bioessays* **21**, 71–75 (1999).
6. Medlin, L. K. et al. *Nova Hewigia* **65**, 1–11 (2000).
7. Round, F. E. & Crawford, R. M. *Proc. R. Soc. Lond. B* **211**, 237–260 (1981).
8. Allison, C. W. & Hilgert, J. W. J. *Paleontol.* **60**, 973–1015 (1986).
9. Tréguer, P. & Pondaven, P. *Nature* **406**, 358–359 (2000).
10. Thomas, D. *Collected Poems of Dylan Thomas 1934–1952* (Norton, New York, 1971).

aids or induces expression of the *Shh* gene in the posterior part of the bud^{1–3}. The three groups reporting in *Development* found that *dHAND* is first expressed throughout the part of the embryo from which limb and fin buds develop (the lateral plate mesoderm; Fig. 1), before limb budding or *Shh* expression begins. As limb buds emerge, *dHAND* expression becomes restricted to the posterior regions of forelimb and hindlimb buds, and to the non-limb-forming region between forelimb and hindlimb buds known as the flank. Only a subset of these *dHAND*-expressing cells will go on to express *Shh*.

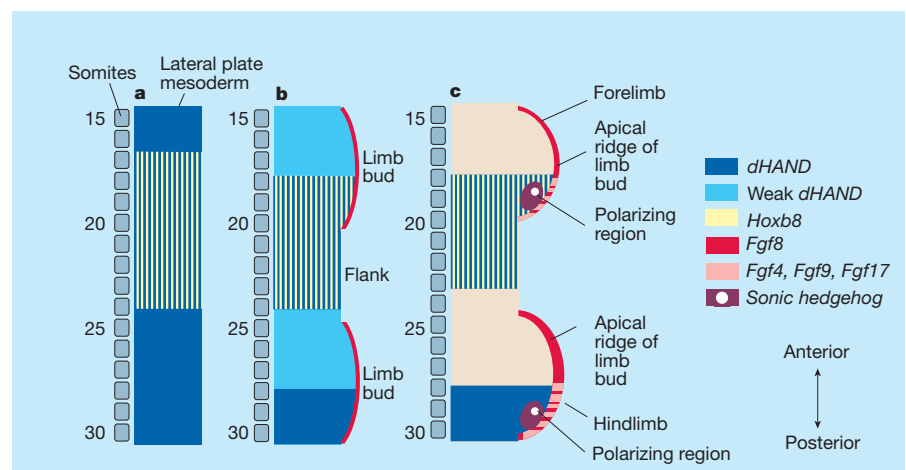


Figure 1 Development of the polarizing region in paired appendages, as suggested by the new work^{1–3}. **a**, Before limb and fin buds begin to form, *dHAND* is expressed throughout the lateral plate mesoderm. Within the *dHAND* territory is an area of *Hoxb8* expression extending from the posterior region of the prospective forelimb bud down to the posterior part of the flank, which corresponds to a region opposite somites (segmented blocks of mesoderm) 17 to 24 in the chick embryo. **b**, As the limb buds start to form, high levels of *dHAND* expression are maintained posteriorly in limb buds and in the flank, but levels decrease in anterior forelimb and hindlimb. **c**, *Sonic hedgehog* expression is induced in limb bud mesenchyme where the domain of *dHAND* expression (and of *Hox8* expression in the forelimb) intersects with the posterior edge of the apical ridge (along the leading edge of the limb), which expresses fibroblast growth factors *Fgf4*, *Fgf8*, *Fgf9* and *Fgf17*. This defines the polarizing region.

Box 1 Evolution of fins and limbs

Four-legged (tetrapod) vertebrates are a subdivision of the lobe-finned fish. As such, forelimbs and hindlimbs are actually modified pectoral and pelvic fins. Most vertebrates have two sets of paired limbs or fins, but none have more than two. These appendages are built along a similar plan and have to solve many of the same problems during development, such as initiation of budding, outgrowth, and patterning the skeleton along three axes.

Although fin development has not been studied in

lobe-finned fish, analyses of fin development in zebrafish (a more distantly related, ray-finned fish) have uncovered remarkable similarities with limb development at the cellular and molecular levels. Some of these characteristics are global features of all animal outgrowths, but others specifically highlight the homology of fins and limbs.

Fin and limb buds are patterned by a posterior polarizing region, which expresses the *Sonic hedgehog* gene, and bud outgrowth is

sustained by fibroblast growth factors secreted by epithelial cells at the distal tip (the apical ridge; Fig. 1). These features reflect the common origin of fins and limbs.

Studies of the transcription factor dHAND in a variety of vertebrates^{1–3} indicate that its role in establishing the polarizing region is also common to fins and limbs. This role of dHAND may be as ancient as paired fins themselves — the earliest known fin skeletons have anterior–posterior asymmetry.

M.J.C.

results in a failure to activate *Shh*, and in the fish causes an absence of paired fins (the mice die shortly after limb budding). At face value, these data suggest a linear genetic cascade of induction, although — as ever — the reality is more complex.

Molecular feedback loops are an integral part of limb development, and several of the new results point to a regulatory feedback loop between dHAND and *Shh*. Overexpression of either dHAND or *Shh* can induce the expression of the other gene. Moreover, mice in which dHAND is knocked out do not express *Shh*, and in *Shh*-knockout mice dHAND expression is severely reduced^{1,2}. *Shh* also participates in a positive feedback loop with signalling molecules known as fibroblast growth factors (FGFs), and these also seem to be needed for dHAND expression². Further experiments should determine whether FGFs regulate *Shh* and dHAND independently.

Understanding the mechanism by which dHAND functions in the vertebrate limb should be helped by analyses of its roles in heart, paired-fin and face development. In zebrafish *Hands off* mutants — which have severe defects of the heart, pectoral fins and jaws — precursors of the heart and pectoral fins are specified but then fail to differentiate. It is not yet known what happens to these cells, but in dHAND-knockout mice, underdevelopment of the branchial arches (from which structures such as jaws and gills form) and the right ventricle of the heart is a result of extensive cell death through apoptosis^{4,5}.

In any developing system that requires a large increase in the number of cells (as in heart, face, fin and limb development), the benefits of proliferation are realized only if the cells are kept alive⁸. If limb budding requires not only cell proliferation, but also active inhibition of apoptosis, then one interpretation of the limb and fin defects in

dHAND mutants is that cells in the limb bud suffer the same apoptotic fate as cells in the face and heart of dHAND-knockout mice.

The results of dHAND overexpression may also be consistent with this interpretation. Mesenchymal cells at the anterior margin of limb buds undergo apoptosis during normal limb development, but cell death is attenuated or absent in several multidigitated mutants⁹. Moreover, anterior limb-bud cells, which normally lack polarizing activity, can develop it under experimental conditions¹⁰. So if the normal function of dHAND is to keep cells with polarizing potential alive in the posterior part of the limb, then overexpression of dHAND might prevent apoptosis in anterior cells, which in turn could induce digit duplications. The ability of *Shh*, retinoic acid and polarizing-region grafts to induce dHAND expression suggests a mechanism by which the polarizing region could ensure its own survival. The situation might be similar in the face, where *Shh* is needed for neural crest cells to survive¹¹. Whether dHAND really does exert its effect on the limb by modulating cell death remains to be tested experimentally. But if true, we may have a new tool for studying the cell biology of embryonic development. ■

Martin J. Cohn is in the School of Animal and Microbial Sciences, University of Reading, Whiteknights, Reading RG6 6AJ, UK.
e-mail: m.j.cohn@reading.ac.uk

- Charité, J., McFadden, D. G. & Olson, E. N. *Development* **127**, 2461–2470 (2000).
- Fernandez-Teran, M. et al. *Development* **127**, 2133–2142 (2000).
- Yelon, D. et al. *Development* **127**, 2573–2582 (2000).
- Srivastava, D. *Trends Cardiovasc. Med.* **9**, 11–18 (1999).
- Thomas, T. et al. *Development* **125**, 3005–3014 (1998).
- Tanaka, M. et al. *Development* (in the press).
- Charité, J. et al. *Cell* **78**, 589–601 (1994).
- Searls, R. L. & Janners, M. Y. *Dev. Biol.* **24**, 198–213 (1971).
- Krabbenhoft, K. M. & Fallon, J. F. *Dev. Dyn.* **194**, 52–62 (1992).
- Anderson, R. et al. *Dev. Biol.* **164**, 241–257 (1994).
- Ahlgren, S. C. & Bronner-Fraser, M. *Curr. Biol.* **9**, 1304–1314 (1999).

Daedalus

Power in the dust

Last week Daedalus devised his 'Stressed Powder', whose tiny ring-particles were needle crystals bent into a highly strained circular form. A suitable shock, rupturing the tense rings, releases their enormous energy density. DREADCO's chemists are now crystallizing many different substances as Stressed Powders, and exploring their properties.

Hard materials such as metal oxides should, if undisturbed, hold their stress indefinitely. Daedalus hopes to exploit such Stressed Powders as long-lasting, non-chemical insecticides. He points out that insects, being so small, are greatly at risk of dehydration. Even a minute crack in their exoskeletal armour allows their vital fluids to evaporate. If the Stressed Powder were scattered around their living space, they would inevitably set off a few grains of it in their wanderings. Each grain would explode like a tiny landmine, firing its crystal fragments at enormous velocity and lethally puncturing the insect's exoskeleton. No nasty chemical residues would remain. To human hands, the powder would be annoyingly prickly, but would pose no serious threat. Like landmines themselves, the powder would remain lethal for years.

Softer materials, such as salt and sugar, would as Stressed Powders slowly creep and relax their tension. But if used promptly they would dissolve in soup or tea with a vigorous release of energy. The resulting drink would be heated almost to boiling point, creating a new 'instant' product. In the same way, many industrial chemicals might be crystallized in stressed form. Like 'nascent' hydrogen, they should be extremely reactive. They would speed up many reactions, and perhaps undergo new ones. Indeed, even an inert Stressed Powder could be chemically useful. In a solvent that crazed or weakened its crystals, it would break up with intense local heat and vigour. It might catalyse many otherwise tricky reactions, both by its energy, as with ultrasound, and by the sudden release of a vast amount of new surface as its particles disintegrate.

For the ultimate in energy density, Daedalus recommends stressed carbon nanotubes. A nanotube bent into a tiny toroid, and thickened by deposition of successive sleeves of carbon, could store far more energy than dynamite. And its explosive disintegration would release a vast amount of active carbon surface. Sadly, even DREADCO's redoubtable chemists can see no way of making Stressed Nanotube Powder.

David Jones

Oceanic respite for wandering albatrosses

Birds taking time off from breeding head for their favourite long-haul destinations.

What oceanic seabirds do outside their breeding periods is something of a mystery, although altogether these 'sabbaticals' add up to more than half of their lifetime and are probably a key feature of their life history. Here we use geolocation systems based on light-intensity measurements to show that during these periods wandering albatrosses (*Diomedea exulans*) leave the foraging grounds that they frequent while breeding for specific, individual oceanic sectors and spend the rest of the year there — each bird probably returns to the same area throughout its life. This discovery of individual home-range preferences outside the breeding season has important implications for the conservation of albatrosses threatened by the development of longline fisheries.

On 26 October 1999, a wandering albatross of at least 50 years old was found dead on a beach in New South Wales, Australia (Fig. 1). A ring identified the bird as coming from the Crozet Islands, 7,900 km away in the Indian Ocean; this bird was originally banded in September 1960 as an adult (over 10 years old) a few kilometres from the site of its death. This recovery of a bird after 40 years, together with the successive recapture every second year of several individuals from Crozet at the same site¹, suggests that wandering albatrosses, which breed every second year, may return routinely throughout their lifetime to a particular place during the non-breeding year.

It has been assumed that these birds wander aimlessly during their biennial sabbatical year, circumnavigating the Southern Ocean by riding the wind^{2,3}, visiting places like the waters of New South Wales en route. The question of where these birds spend half of their life is important (Fig. 2), not only because this species may potentially exploit almost any area in the Southern Ocean⁴, but also because wandering albatrosses, like many other albatrosses and petrels, are threatened by longline fisheries in the Southern Ocean. They are killed in tens of thousands while scavenging for baited hooks and from drowning when lines are set⁵. Any plan to remedy this shocking loss will depend on establishing the extent of overlap in the areas used by both albatrosses and fisheries.

Using satellite telemetry to study the movements of seabirds over a complete year is difficult and costly. Instead, we determined bird position using loggers weighing 20 g that record light intensity and are attached to a ring on the bird's leg. We were able to estimate the local time of dawn and

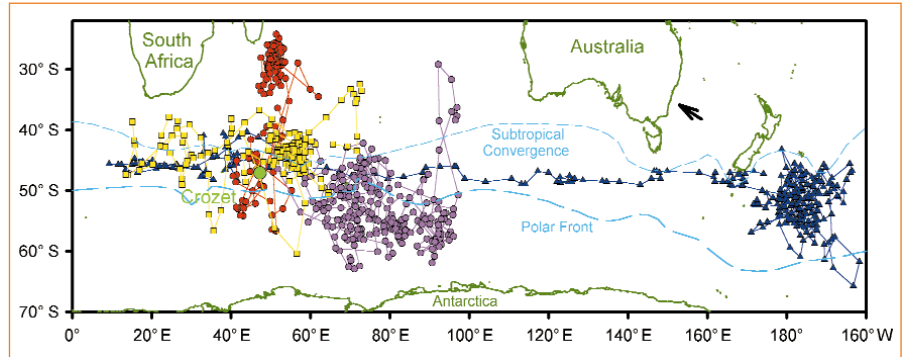


Figure 1 Oceanic sectors used by two male (mauve dots and blue triangles) and two female (red dots and yellow squares) wandering albatrosses, as studied by geolocation. The dots indicate two partners of the same pair. Two locations are estimated for each bird per 24-h cycle. Arrow indicates the recovery site of ringed birds off New South Wales, Australia; green circle, the Crozet Islands.

dusk, determined by light thresholds, with respect to Greenwich Mean Time using a precise internal clock: after consideration of the date, the day length could be used to determine the latitude, and the longitude was determined from local time of noon^{6–8}.

We fitted nine wandering albatrosses on the Crozet Islands with loggers during February–March in 1995 and 1996, just after a failed breeding attempt and before their departure from the breeding grounds for a year. This pattern of absence from the breeding grounds is similar to that adopted by birds that complete a successful breeding season in December¹. All tagged birds returned one year later with the unit intact, although only four devices provided light measurements over 6 months or more.

After leaving the island, the four birds (two males and two females) flew to a specific ocean sector where they overwintered, returning to their breeding place in late December. These sectors ranged from tropical and subtropical waters (the females) to sub-Antarctic and Antarctic waters (the males) at distances 1,500–8,500 km away from Crozet. In each pair, the male spent the winter just north of the pack ice in Antarctic waters, whereas the female stayed south of Madagascar.

To our knowledge, this is the first time that the wintering zones of a pelagic seabird have been studied remotely. Our results provide surprising evidence that wandering albatrosses do not wander around the Southern Ocean during their non-breeding year. After spending a year covering an estimated 150,000 km foraging around colonies to hatch and raise their single chick^{4,9}, wandering albatrosses take a year off. The vast distances covered during breeding increase the probability of encountering dispersed prey while effectively using the wind to reduce flight costs⁹.

During the sabbatical year, the birds have lower energy requirements and are not restricted to a central breeding place.

But even with this new freedom, the birds confine their movements to a preferred sector, not only because of their reduced energy requirements but maybe also because at this time they undergo a partial moult of their flight feathers. Moulting probably reduces flight efficiency and therefore increases flight costs. Also, by moving away to a distant ocean sector far from the breeding grounds, the competition between the element of the population in its breeding year and the other in its sabbatical year is reduced. Information from banding indicates that birds use the same wintering zone from year to year, probably throughout their lifetime. The wintering grounds favoured by each individual in the open ocean may be visited repeatedly because the birds learn over years the location of profitable feeding zones.

Our results have important implications



Figure 2 The wandering albatross spends more than 95% of its life in the open ocean, ten years of this as an immature bird, but also half of its mature life during its non-breeding years. Birds move to specific oceanic areas far from their breeding grounds; area preferences appear to last a lifetime.

for the conservation of this threatened species¹⁰. During the breeding season, females favour the subtropical waters north of Crozet where tuna fisheries are located, whereas males prefer the colder waters at higher latitudes where fisheries for toothfish have been developed¹¹. Birds spending their sabbatical year in sectors where long-line fisheries occur are at risk of being killed. This bizarre selection pressure, dependent perhaps on the whims of juvenile choice, means that only those wintering in zones without fisheries will survive in the long term. We now need to find out what factors lead to selection of sabbatical areas in the first place. In the meantime, we have to accept that even during their sabbaticals adult birds are unlikely to change their habits to avoid the dangers from fisheries.

Henri Weimerskirch*†, Rory P. Wilson‡

*CEBC-CNRS, 79360 Villiers en Bois, France
e-mail: henriw@cebc.cnrs.fr

†Institut Français pour la Recherche et la

Technologie Polaire, 29280 Plouzané, France

‡Institut für Meereskunde, D-24105 Kiel, Germany

1. Weimerskirch, H., Jouventin, P., Mougou, J. L., Stahl, J. C. & Van Beveren, M. *Emu* **85**, 22–23 (1985).
2. Murphy, R. C. *Oceanic Birds of South America* (Macmillan, New York, 1936).
3. Nicholls, D. G., Murray, M. D., Butcher, E. & Moors, P. *Emu* **97**, 240–244 (1997).
4. Jouventin, P. & Weimerskirch, H. *Nature* **343**, 746–748 (1990).
5. Brothers, N. P. *Biol. Conserv.* **55**, 255–268 (1991).
6. Wilson, R. P., Ducamp, J. J., Rees, G., Culik, B. M. & Niekamp, K. in *Wildlife Telemetry: Remote Monitoring and Tracking of Animals* (eds Priede, I. M. & Swift, S. M.) 131–134 (Horward, Chichester, 1992).
7. Hill, R. D. in *Elephant Seals: Population Ecology, Behavior, and Physiology* (eds Le Boeuf, B. J. & Laws, R. M.) 227–236 (Univ. California Press, Berkeley, 1994).
8. Welch, D. W. & Eveson, J. P. *Can. J. Fish. Aquat. Sci.* **56**, 1317–1327 (1999).
9. Weimerskirch, H., Doncaster, P. & Cuénot-Chaillet, F. *Proc. R. Soc. Lond. B* **255**, 91–97 (1994).
10. Robertson, G. & Gales, R. *Albatross, Biology and Conservation* (Beatty, Chipping Norton, Australia, 1993).
11. Weimerskirch, H., Brothers, N. P. & Jouventin, P. *Biol. Conserv.* **79**, 257–270 (1997).

Plant biotechnology

Caffeine synthase gene from tea leaves

Caffeine synthase is an enzyme that catalyses the final two steps in the caffeine biosynthesis pathway. We have cloned the gene encoding caffeine synthase from young leaves of tea (*Camellia sinensis*), opening up the possibility of creating tea and coffee (*Coffea arabica*) plants that are naturally deficient in caffeine. Consumers concerned about the possible adverse effects of caffeine consumption will welcome this development towards caffeine-free drinks that retain their flavour.

The increasing demand for decaffeinated coffee and tea has resulted from the occasional side effects of caffeine, which include palpitations, gastrointestinal disturbances, anxiety, tremor, increased blood pressure and insomnia^{1,2}. At present, the decaffeination process depends on supercritical fluid extraction with carbon dioxide to avoid introducing toxic residues from extraction solvents. However, this operation is expensive and, to discerning customers, flavours and aromas are lost.

There are some *Coffea* and *Camellia* species that produce low levels of caffeine, but these are not readily available for commercial use. Although it might be feasible to develop a breeding programme, it could take 20 years or so to establish and stabilize the desired traits³. The large-scale production of transgenic caffeine-deficient *Camellia sinensis* and *Coffea arabica* plants may thus be a more practical proposition⁴, but first some information is needed about the genes controlling key conversions in the biosynthesis of caffeine.

Caffeine is synthesized from purine nucleotides. The two final steps of the pathway, in which two methyl groups are added successively to 7-methylxanthine to produce theobromine and then caffeine, are catalysed by caffeine synthase, a bifunctional enzyme comprising two S-adenosylmethionine-dependent N-methyltransferase activities⁵. It has been hard to purify and isolate caffeine synthase and other enzymes of this pathway because they are extremely labile, but the amino-terminal sequence of caffeine synthase from young tea leaves has now been reported⁵.

We used the RACE (rapid amplification of complementary DNA ends) technique with degenerate gene-specific primers based on the amino-terminal sequence of caffeine synthase to obtain a 1.31-kilobase sequence of cDNA. The 5' untranslated sequence of the cDNA fragment was isolated by 5' RACE. The isolated cDNA, termed TCS1 (GenBank accession no. AB031280), consists of 1,438 base pairs and encodes a protein of 369 amino acids. The deduced

amino-acid sequence of TCS1 shares a small amount of sequence similarity with other N-, S- and O-methyltransferases from plants and microorganisms, but considerably more with the salicylic acid O-methyltransferase⁶ (41.2%).

To determine whether our cDNA encoded an active caffeine synthase enzyme, we expressed TCS1 in *Escherichia coli* and incubated lysates of the bacterial cells with a variety of xanthine substrates in the presence of S-adenosylmethionine as methyl donor. We found that the substrate specificity of the recombinant enzyme was very similar to that of the native enzyme purified from young tea leaves (Table 1), with the recombinant enzyme mainly catalysing 3-N-methylation and 1-N-methylation of the purine ring of mono- and dimethylxanthines. There was no 7-N-methylation activity when xanthosine was the methyl acceptor. These results indicate that TCS1 encodes caffeine synthase.

The cloning of the caffeine synthase gene is an important advance towards the development of transgenic caffeine-deficient *Camellia sinensis* and *Coffea arabica* plants through antisense messenger RNA technology or by gene silencing. It is possible that the health benefits of tea^{7–11}, whose catechins and related polyphenols^{7–10} are thought to help protect against heart disease, may be enhanced without the potentially hypertensive effects of caffeine^{12,13}.

Misako Kato*, Kouichi Mizuno†, Alan Crozier‡, Tatsuhito Fujimura†, Hiroshi Ashihara*

*Department of Biology, Faculty of Science, Ochanomizu University, Otsuka, Bunkyo-ku, Tokyo 112-8610, Japan

†Institute of Agricultural and Forest Engineering, University of Tsukuba, Ibaraki 305-8572, Japan

‡Plant Products and Human Nutrition Group, Division of Biochemistry and Molecular Biology, Institute of Biomedical and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK
e-mail: a.crozier@bio.gla.ac.uk

1. Chou, T. M. & Benowitz, N. L. *Comp. Biochem. Physiol.* **109C**, 173–189 (1994).
2. Nurminen, M.-L., Niittynen, L., Korpela, R. & Vapaatalo, H.

Table 1 Substrate specificity of recombinant and native caffeine synthase

Substrate	Methylated product	N-methylation position	Recombinant CS	Native CS*
Monomethylxanthines				
7-Methylxanthine	Theobromine	3	100†	100
3-Methylxanthine	Theophylline	1	1.0	17.6
1-Methylxanthine	Theophylline	3	12.3	4.2
Dimethylxanthines				
Theobromine	Caffeine	1	18.5	26.8
Theophylline	Caffeine	7	<0.1	<0.1
Paraxanthine	Caffeine	3	230	210
Others				
Xanthosine	7-Methylxanthosine	7	Not detected	Not detected

The full-length coding region for the tea-leaf caffeine synthase (CS) protein was ligated into pET23d plasmids and the resultant expression vector introduced into *E. coli* (BL21). Recombinant CS protein was extracted by sonication of the transformed cells in 50 mM Tris-HCl (pH 7.5) containing 1 mM EDTA-Na₂ and 0.1 M NaCl. The substrate specificity of recombinant CS was determined according to ref. 5. CS activity is expressed as a percentage of the activity on 7-methylxanthine.

*Caffeine synthase activity of the recombinant enzyme with 7-methylxanthine (100%) was 5.4 pkat mg⁻¹ protein; values represent the average of duplicate samples.

†Taken from ref. 5.

- Eur. J. Clin. Nutr. 53, 831–839 (1999).
3. Ashihara, H. & Crozier, A. *Adv. Bot. Res.* 30, 117–205 (1999).
 4. Mazzafera, P., Crozier, A. & Magalhães, A. C. *Phytochemistry* 30, 3913–3916 (1991).
 5. Kato, M. *et al. Plant Physiol.* 120, 579–586 (1999).
 6. Ross, J. R., Nam, K. H., D'Auria, J. C. & Pichersky, E. *Arch. Biochem. Biophys.* 367, 9–16 (1999).
 7. Stensvold, I., Tverdal, A., Solvoll, K. & Foss, O. P. *Prev. Med.* 21, 546–553 (1992).
 8. Hertog, M. G. L. *et al. Lancet* 342, 1007–1010 (1993).
 9. Vinson, J. A. & Dabbagh, Y. F. *FEBS Lett.* 433, 44–46 (1998).
 10. Leenen, R., Roodenburg, A. J. C., Tijburg, L. B. M. & Wiseman, S. A. *Eur. J. Clin. Nutr.* 54, 87–92 (2000).
 11. Fujiki, H. J. *Cancer Res. Clin. Oncol.* 125, 589–597 (1999).
 12. Lane, J. D. *et al. Psychosom. Med.* 60, 327–330 (1998).
 13. James, J. E. *Lancet* 349, 279–281 (1997).

Neurobiology

Kainate receptors and synaptic plasticity

Bortolotto *et al.*¹ report that the kainate subtype of glutamate receptor is essential for the plasticity of certain types of synaptic transmission in the brain, which is of interest as these receptors were previously not thought to initiate plastic processes. In particular, a new antagonist (LY382884) was shown to act selectively against the GluR5 type of kainate receptor: in the presence of LY382884, which reduces kainate-receptor-mediated postsynaptic responses by ~40%, long-term potentiation (LTP) at hippocampal mossy-fibre synapses could no longer be induced¹. Here we argue that the available evidence does not support a major role for kainate receptors in the induction of mossy-fibre LTP.

It is well established that kainate receptors are also blocked by the ionotropic glutamate-receptor antagonists kynurenate and CNQX, and therefore, if Bortolotto *et al.* are correct, mossy-fibre LTP should also be blocked by these antagonists. Although they do provide evidence that this is the case, their results contradict earlier work showing that these antagonists, at the same or higher concentrations, are ineffective on mossy-fibre LTP^{2–6}.

In two of these studies^{2,3}, 10–20 mM kynurenate was found to have no effect, whereas Bortolotto *et al.*¹ report that 10 mM kynurenate blocks LTP completely (for reference, 1 mM kynurenate is sufficient to block the kainate-receptor-mediated excitatory postsynaptic current (e.p.s.c.)⁷). The others^{4–6} showed that 10–20 μ M CNQX, which also blocks the kainate-receptor-mediated e.p.s.c. by 40–80% (ref. 8), did not prevent induction of mossy-fibre LTP. In two cases^{5,6}, the entire experiment was done in the presence of 10 μ M CNQX, the concentration used to block mossy-fibre LTP by Bortolotto *et al.*¹, and the increase in glutamate release that underlies mossy-fibre LTP was monitored using the NMDA-receptor-mediated e.p.s.c.

By studying LTP on the NMDA-receptor component of the e.p.s.c., which has properties identical to the LTP of the AMPA-receptor component^{5,6}, the potentiation can be assessed continuously after induction. We believe that this approach has better resolving power than the experiments of Bortolotto *et al.*¹, where LTP is presumed to be blocked on the basis that no potentiation is apparent three hours after induction and washout of the antagonist. Evidence using glutamate-receptor antagonists^{2–6} would indicate that kainate receptors are not required for mossy-fibre LTP.

The results taken together suggest that the action of LY382884 on mossy-fibre LTP is either nonspecific or indirect. This is supported by expression data showing that the GluR5 subunit of kainate receptors is only weakly expressed in the dentate gyrus and in area CA3 of the hippocampus^{9,10}, and that most (about 90%) of the cells generating this weak signal are GABAergic interneurons¹⁰, thought not to be directly involved in the induction of mossy-fibre LTP.

Given these combined results^{2–6,9,10}, we believe that GluR5 is unlikely to play a role in mossy-fibre synaptic transmission or plasticity; although LY382884 may antagonize GluR5-containing kainate receptors, its effect on mossy-fibre LTP is more likely to involve some other interaction of this drug with synaptic transmission.

R. A. Nicoll, J. Mellor, M. Frerking, D. Schmitz

Department of Cellular and Molecular Pharmacology, University of California, San Francisco, California 94143-0450, USA
e-mail: nicoll@phy.ucsf.edu

1. Bortolotto, Z. A. *et al. Nature* 402, 297–301 (1999).
2. Castillo, P. E., Weisskopf, M. G. & Nicoll, R. A. *Neuron* 12, 261–269 (1994).
3. Yeckel, M. F., Kapur, A. & Johnston, D. *Nature Neurosci.* 2, 625–633 (1999).
4. Ito, I. & Sugiyama, H. *NeuroReport* 2, 333–336 (1991).
5. Weisskopf, M. G. & Nicoll, R. A. *Nature* 376, 256–259 (1995).
6. Tong, G., Malenka, R. C. & Nicoll, R. A. *Neuron* 16, 1147–1157 (1996).
7. Castillo, P. E., Malenka, R. C. & Nicoll, R. A. *Nature* 388, 182–186 (1997).
8. Vignes, M. & Collingridge, G. L. *Nature* 388, 179–182 (1997).
9. Bahn, S., Volk, B. & Wisden, W. *J. Neurosci.* 14, 5525–5547 (1994).
10. Paternain, A. V., Herrera, M. T., Nieto, M. & Lerma, J. *J. Neurosci.* 20, 196–205 (2000).

Bortolotto et al. reply — Nicoll *et al.* challenge our finding that kainate receptors are involved in mossy-fibre LTP in the hippocampus¹. Their argument is based on a discrepancy of our results with earlier work, as we show (our Fig. 5)¹ that two kainate-receptor antagonists, kynurenate and CNQX, can also block mossy-fibre LTP. We do not dispute that it may be possible to induce mossy-fibre LTP in the presence of kainate-receptor antagonists under certain conditions. For this reason, we neither stated nor implied that activation of

kainate receptors is always necessary for mossy-fibre LTP¹.

Concerning the resolving power of our experiments, it is important to consider three factors. First, although the effect of CNQX on the induction of LTP had to be assessed three hours after the tetanus owing to the slow washout of CNQX, we often included a non-tetanzed input as a control, which also invariably returned to the same, pre-CNQX, level². Second, in our experiments with the selective AMPA-receptor antagonist GYKI53655, we also had to measure LTP three hours after washout of the antagonist, and LTP was always observed in the tetanzed input². Finally, the kynurenate experiments required only one hour for complete washout.

The relatively low expression of GluR5 messenger RNA in principal cells within the dentate gyrus and area CA3 was not a major concern for us because the relation between GluR5 gene expression and protein levels at this synapse is not known. In addition, as discussed previously³, it is possible that the LY382884 class of compounds can also act on GluR5-lacking heteromeric assemblies of kainate receptors, such as GluR6-KA1, which have not been tested with these antagonists.

The fact that others have observed mossy-fibre LTP in the presence of kynurenate and CNQX has several possible explanations, the most likely being, we suspect, that involvement of GluR5-containing kainate receptors in mossy-fibre LTP can be bypassed in certain circumstances. To determine conditions under which LY382884 might fail to block the induction of mossy-fibre LTP, we have applied multiple trains at test intensity and find that LTP is always blocked by the antagonist; however, when we greatly increase the stimulus strength during the tetanus, some LTP remains².

Multiple routes for the induction of LTP at CA1 synapses, involving both NMDA⁴ and mGlu⁵ receptors, have already been reported. Crosstalk between receptors and their signalling cascades may make LTP more difficult to understand, but it provides synapses with a much richer array of mechanisms with which to build memories. **Zuner A. Bortolotto***, **Vernon R. J. Clarke***, **Caroline M. Delany***, **Michel Vignes†**, **Graham L. Collingridge***

*MRC Centre for Synaptic Plasticity, Department of Anatomy, Medical School, University of Bristol, Bristol BS8 1TD, UK
e-mail: g.l.collingridge@bristol.ac.uk

†Laboratoire 'Plasticité Cérébrale', EP 628 CNRS, Université Montpellier II, Place Eugène Bataillon, 34095 Montpellier Cedex 05, France

1. Bortolotto, Z. A. *et al. Nature* 402, 297–301 (1999).
2. www.bris.ac.uk/synaptic/info/kainate.html
3. Vignes, M. *et al. Neuropharmacology* 37, 1269–1277 (1998).
4. Grover, L. M. & Teyler, T. J. *Nature* 347, 477–479 (1990).
5. Bortolotto, Z. A., Bashir, Z. I., Davies, C. H. & Collingridge, G. L. *Nature* 368, 740–743 (1994).

Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen

C. K. Stover*, X. Q. Pham†, A. L. Erwin*, S. D. Mizoguchi*, P. Warren*, M. J. Hickey*, F. S. L. Brinkman‡, W. O. Hufnagle*, D. J. Kowalik*, M. Lagrou*, R. L. Garber*, L. Goltry*, E. Tolentino*, S. Westbrook-Wadman*, Y. Yuan*, L. L. Brody*, S. N. Coulter*, K. R. Folger*, A. Kas†, K. Larbig§, R. Lim†, K. Smith†, D. Spencer†, G. K.-S. Wong†, Z. Wu†, I. T. Paulsen||, J. Reizer¶, M. H. Saier¶, R. E. W. Hancock‡, S. Lory# & M. V. Olson†

* PathoGenesis Corporation, 201 Elliott Avenue West, Seattle, Washington 98119, USA

† Department of Medicine and Genetics, University of Washington Genome Center, Box 352145, University of Washington, Seattle, Washington 98195, USA

‡ Department of Microbiology and Immunology, University of British Columbia, 300-6174 University Blvd, Vancouver, British Columbia V6T 1Z3, Canada

§ Klinische Forschergruppe, Medizinische Hochschule Hannover, D-30623 Hannover, Germany

|| The Institute for Genomic Research, 9712 Medical Center, Rockville, Maryland 20850, USA

¶ Department of Biology, University of California at San Diego, 9500 Gilman Drive, La Jolla, California 92093-0116, USA

Department of Microbiology, University of Washington School of Medicine, Seattle, Washington 98195, USA

***Pseudomonas aeruginosa* is a ubiquitous environmental bacterium that is one of the top three causes of opportunistic human infections. A major factor in its prominence as a pathogen is its intrinsic resistance to antibiotics and disinfectants. Here we report the complete sequence of *P. aeruginosa* strain PA01. At 6.3 million base pairs, this is the largest bacterial genome sequenced, and the sequence provides insights into the basis of the versatility and intrinsic drug resistance of *P. aeruginosa*. Consistent with its larger genome size and environmental adaptability, *P. aeruginosa* contains the highest proportion of regulatory genes observed for a bacterial genome and a large number of genes involved in the catabolism, transport and efflux of organic compounds as well as four potential chemotaxis systems. We propose that the size and complexity of the *P. aeruginosa* genome reflect an evolutionary adaptation permitting it to thrive in diverse environments and resist the effects of a variety of antimicrobial substances.**

P. aeruginosa is a versatile Gram-negative bacterium that grows in soil, marshes and coastal marine habitats, as well as on plant and animal tissues¹. It forms biofilms on wet surfaces such as those of rocks and soil^{2,3}. The emergence of *P. aeruginosa* as a major opportunistic human pathogen during the past century may be a consequence of its resistance to the antibiotics and disinfectants that eliminate other environmental bacteria. *P. aeruginosa* is now a significant source of bacteraemia in burn victims, urinary-tract infections in catheterized patients, and hospital-acquired pneumonia in patients on respirators⁴. It is also the predominant cause of morbidity and mortality in cystic fibrosis patients, whose abnormal airway epithelia allow long-term colonization of the lungs by *P. aeruginosa*. These infections are impossible to eradicate, in part because of the natural resistance of the bacterium to antibiotics, and ultimately lead to pulmonary failure and death.

Here we report the sequencing of the genome of *P. aeruginosa*. The sequence is of interest because of the insights it provides into the role of this bacterium as a pathogen, and because it offers new information on the relationship between genome size, genetic complexity and ecological versatility in bacteria. At 6.3 million base pairs (Mbp), the *P. aeruginosa* genome is markedly larger than most of the 25 sequenced bacterial genomes. In fact, with 5,570 predicted open reading frames (ORFs), the genetic complexity of *P. aeruginosa* approaches that of the simple eukaryote *Saccharomyces cerevisiae*, whose genome encodes about 6,200 proteins⁵. In contrast, *P. aeruginosa* has only 30–40% of the number of predicted genes present in the simple metazoans *Caenorhabditis elegans* and *Drosophila melanogaster*⁶.

genome-shotgun sampling. The largest genome that has been completely sequenced by this approach is from *Deinococcus radiodurans* (2.6 Mbp)⁷. The other large bacterial genome sequences (*Bacillus subtilis*, 4.2 Mbp; *Synechocystis*, 3.6 Mbp; *Escherichia coli*, 4.6 Mbp; and *Mycobacterium tuberculosis*, 4.4 Mbp)^{8–11} were all initially determined by sequencing overlapping sets of clones, polymerase chain reaction (PCR) products and gel-purified restriction fragments.

With one major exception, the assembled genome sequence is in excellent agreement with the physical map of the *P. aeruginosa* genome^{12,13}. The exception is the inversion of more than one-quarter of the genome in the PA01 isolate we sequenced, relative to DSM-1707, the PA01-derived isolate previously mapped in the laboratory of B. Tümmler^{12,13} (Fig. 1). As both of these isolates are clonally derived from the original PA01 strain of *P. aeruginosa*, any differences in genome structure between PA01 and DSM-1707 must have arisen during propagation. This inversion does not appear to be unique to the sequenced isolate as PA01 stocks from other laboratories have the same inversion (H. Schweizer, personal communication). The inversion appears to have resulted from homologous recombination between the *rrnA* and *rrnB* loci, which are orientated in opposite directions and separated by 1.7 Mbp (Fig. 1). Comparative analysis of digests of PA01 and DSM-1707 with *SfoI*, *SwaI* and *PacI* supported this possibility. Earlier observations of inversions of genomic segments between oppositely orientated ribosomal DNA loci in *E. coli* and *S. typhimurium* led to the proposal that these reversible genome rearrangements may have adaptive significance¹⁴.

Sequencing and assembly

Sequencing of the complete 6.3-Mbp genome of *P. aeruginosa* was accomplished by a straightforward implementation of whole-

Properties of the genome and relationship to other bacteria

Basic features. These are summarized in Table 1 and Fig. 1. Most of the predicted ORFs have the high G+C content (66.6%)

characteristic of the genome as a whole and have codon usage similar to previously described *P. aeruginosa* genes. However, ten regions of 3.0 kilobases (kb) or greater exhibit significantly lower G+C content and unusual codon usage (Fig. 1), possibly indicative of recent horizontal transfer. In addition, there are two regions (PA616–PA648, PA715–PA728) containing probable bacteriophages.

Comparative analysis. To gain insight into the significance of the size of this genome, we concentrated on comparative analysis of the *P. aeruginosa* and *E. coli* genomes. Not only is *E. coli* the most intensively studied of all bacteria, but it is also the closest relative of *P. aeruginosa* among the bacteria with fully sequenced genomes: for example, when each of the 5,570 predicted ORFs for *P. aeruginosa* was compared with the pooled ORFs for 22 other bacterial genomes by the sequence-alignment algorithm BLASTP⁴⁴, nearly half of the best hits above a stringent comparison threshold (an expect value of 10^{-5}) were to *E. coli*, and no other organism accounts for even 10% (see Supplementary Information). The relative prominence of *E. coli* increases moderately at more stringent thresholds although a noisy pattern of weak 'best hits' to phylogenetically distant bacteria emerges at lower thresholds. Although this test confirms that *E. coli* is a sensible comparison partner for *P. aeruginosa*, the median amino-acid identity within the aligned region of the *P. aeruginosa*–*E. coli* orthologues is only 40%.

Comparison of the *P. aeruginosa* and *E. coli* genomes indicates that the large genome of *P. aeruginosa* is the result of greater genetic complexity rather than differences in genome organization. Distributions of ORF sizes and inter-ORF spacings are both nearly identical in the two genomes (see <http://www.pseudomonas.com>), and the extent of evolutionarily recent duplications appears comparable. The longest repeats in the *P. aeruginosa* genome are the four rDNA loci and one duplicated gene cluster that spans a few thousand base pairs (PA1899–PA1905; PA4210–PA4216). At the level of amino-acid sequence conservation, residual stretches of locally conserved gene order between *P. aeruginosa* and *E. coli* are far more evident than are internal duplications in either genome. At the same BLASTP threshold employed for the comparisons between the *P. aeruginosa* ORFs and those of other bacterial genomes, we searched for clusters of five or more ORFs that are conserved between *P. aeruginosa* and *E. coli*, allowing single-ORF insertions or deletions within the clusters. Thirty-three distinct clusters were identified, which included 256 ORFs; seven of these clusters involve ten or more ORFs. This analysis showed only a few gene clusters duplicated within either the *E. coli* or the *P. aeruginosa* genome. Hence, with respect to local gene order, evidence of the common ancestry of segments of the *P. aeruginosa* and *E. coli* genome is far more abundant than are the vestiges of more recent duplication events of comparable size within either genome.

Evidence for increased functional diversity. The apparent lack of recent gene duplication indicates that the size of the *P. aeruginosa* genome is due to greater gene and functional diversity. When we analysed the *P. aeruginosa*, *E. coli*, *B. subtilis* and *M. tuberculosis* genomes by BLASTP comparisons between all predicted ORFs within each organism, we found that the *P. aeruginosa* genome has significantly more distinct gene families (paralogous groups) than the other large bacterial genomes (see Supplementary Information). There are nearly 50% more paralogous groups in *P. aeruginosa* than predicted on the basis of a simple comparison of genome sizes.

If the larger genome of *P. aeruginosa* arose by recent gene duplication, we would have expected it to have a similar number of paralogous groups to the other large bacterial genomes, with a larger number of ORFs in each group. In fact, the number of ORFs in the paralogous groups in PAO1 is similar to the other genomes. These data indicate that selection for environmental versatility has favoured expansion of genetic capability through the development

of numerous small paralogous gene families whose members encode distinct functions.

Annotation of *P. aeruginosa* genes

Prediction of gene function. The predicted ORFs were examined individually for (1) identity with known genes of *P. aeruginosa* with sequences deposited in GenBank, (2) similarity with well-characterized genes from other bacteria, or (3) presence of known functional motifs (Table 1; see <http://www.pseudomonas.com> for complete list). In each case the literature was searched to ensure that the proteins encoded by the homologous genes were functionally characterized to avoid the perpetuation of poorly supported functional assignments. In addition, 61 researchers who were members of the *P. aeruginosa* research community or had experience in particular aspects of bacterial physiology were enlisted for the *Pseudomonas* Community Annotation Project (PseudoCAP) to provide expert assistance and confirmatory information for the identification of ORFs and assigned functions.

We were able to assign a functional class to 54.2% of ORFs (Table 2). As in other bacterial genomes, a large proportion of the genome (45.8% of ORFs) consists of genes for which no function could be determined or proposed (confidence level 4; see Table 1). Of these, nearly a third (769 ORFs) possess homology to genes of unknown function predicted in other bacterial genomes, and the remainder (32% of ORFs) does not have strong homology with any reported sequence.

The 372 ORFs that are known *P. aeruginosa* genes with demonstrated functions (confidence level 1) are primarily genes encoding lipopolysaccharide biosynthetic enzymes, virulence factors, such as exoenzymes and the systems that secrete them,

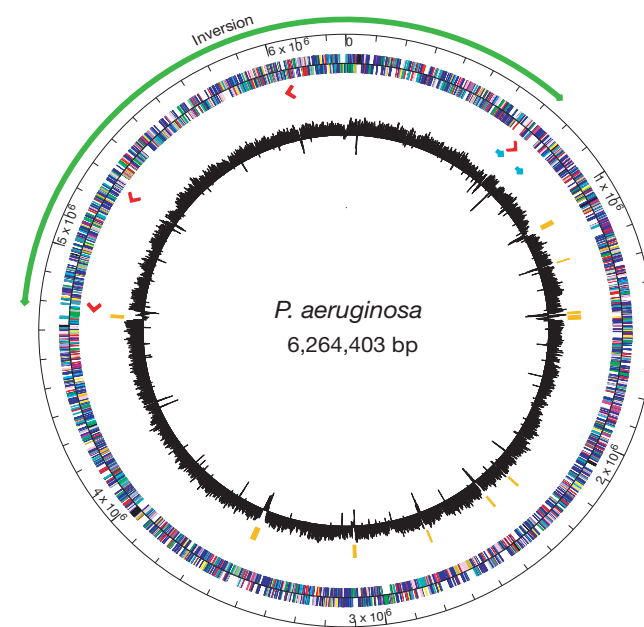


Figure 1 Circular representation of the *P. aeruginosa* genome. The outermost circle indicates the chromosomal location in base pairs (each tick is 100 kb). The distribution of genes is depicted by coloured boxes according to functional category and direction of transcription (outer band is the plus strand; inner band is the minus strand). Red arrows, the locations and direction of transcription of ribosomal RNA genes; green arrow, the inverted region that resulted from a homologous recombination event between *rrmA* and *rrmB*; blue arrows, location of two regions containing probable bacteriophages. The black plot in the centre is percentage G+C content plotted as the average for non-overlapping 1-kb windows spanning one strand for the entire *P. aeruginosa* genome. Yellow bars, regions of ≥ 3.0 kb with G+C content of two standard deviations ($< 58.8\%$) below the mean (66.6%) (see Supplementary Information). A linear map of the genes, with the colour code for functional categories, is available at <http://www.pseudomonas.com>.

and proteins involved in motility and adhesion. ORFs with strong homology to genes in other organisms with demonstrated functions (confidence level 2; 1,059 ORFs) include those required for DNA replication, protein synthesis, cell-wall biosynthesis and intermediary metabolism. *P. aeruginosa* is able to grow on minimal medium, and as we expected, we identified most of the genes required for biosynthesis of amino acids, nucleic acids and cofactors.

The ORFs that provided the most new information about *P. aeruginosa* biology are those that could be assigned a probable function on the basis of similarity to established sequence motifs, but could not be assigned a definite name (confidence level 3; 1,590 ORFs). Most of these genes encode products that are in one of three functional classes: putative enzymes (405 genes), transcriptional regulators (341 genes) or transporters of small molecules (408 genes). In some cases genomic context provided additional information, allowing us to identify loci that appear to encode systems such as metabolic pathways and secretion systems, although the substrates for such systems could not be identified. These and other features of the *P. aeruginosa* genome that may shed light on its biology are discussed below. Additional details are available in the Supplementary Information and at <http://www.pseudomonas.com>.

Regulation. *P. aeruginosa* has the highest proportion of predicted regulatory genes observed in the sequenced bacterial genomes. Analysis using relevant Pfam 5.2 family models¹⁵ and HMMER 2.1.1 (<http://hmmer.wustl.edu/>) shows 468 genes containing motifs characteristic of transcriptional regulators or environmental sensors (see Supplementary Information). This analysis predicts that 8.4% of *P. aeruginosa* genes are involved in regulation, a far higher proportion than is found in other sequenced genomes. (Manual annotation of the genome identified 521 genes (9.4%) as encoding either transcriptional regulators or two-component regulatory system proteins (Table 2). Thus the parameters we employed gave somewhat conservative predictions.)

Similar computational analysis of regulatory motifs in 22 genomes indicates that as bacterial genome size increases, the proportion of the genome devoted to regulatory proteins increases as well (Fig. 2). This trend appears most prominent in prototrophic bacteria that can survive in diverse environments. For example, motifs characteristic of regulatory proteins are found in 5.8% of *E. coli* genes and 5.3% of *B. subtilis* genes, but only in 3.0% of genes

in *M. tuberculosis*, a highly specialized pathogen with a comparable genome size. *Helicobacter pylori*, another highly specialized bacterial pathogen with a much smaller genome, possesses even less regulatory potential (1.1% of genes). When we compared *P. aeruginosa* transcriptional regulators with other bacterial systems, the most striking over-representations were in the LysR, AraC, ECF- σ and two-component regulator families. There is an extraordinary number of putative two-component regulatory system proteins, with 55 sensors, 89 response regulators and 14 sensor-response regulator hybrids, far more than found in the other genomes analysed. Such systems permit organisms to respond to changes in their environment, and are often associated with global regulatory systems as well as with regulation of virulence.

Outer membrane proteins. Outer membrane proteins (OMPs) are of particular interest in *P. aeruginosa* due to their cell-surface exposure and their involvement in transport of antibiotics, in export of extracellular virulence factors, and in anchoring the structures that mediate adhesion and motility. About 150 genes are predicted to encode OMPs, a disproportionately large number compared with other genomes. Three large paralogous families were identified: the OprD family of specific porins (19 genes), the TonB-family of gated porins, which includes proteins involved in iron-siderophore uptake (34 genes), and the OprM family of outer membrane proteins involved in efflux or secretion (18 genes). These large families of proteins were unexpected, as single members of these families (for example, OprD) had been well studied with no appreciation that these proteins were members of a large paralogous group. To date, the only other genome that is known to contain a large paralogous family of OMPs is *H. pylori*¹⁶. The identification of these families could have a significant impact on the focus of antimicrobial and vaccine research.

Import of nutrients. Consistent with its environmental versatility, *P. aeruginosa* has nearly 300 cytoplasmic membrane transport systems, about two-thirds of which appear to be involved in the import of nutrients and other molecules (<http://www-biology.ucsd.edu/~ipaulsen/transport>). The overall substrate specificities of the *P. aeruginosa* transporters are similar to those of *E. coli* and *B. subtilis* with certain significant exceptions (see Supplementary Information). *P. aeruginosa* has a large variety of transporters for mono-, di-, and tri-carboxylates, but it appears to be conspicuously deficient in sugar transporters. For example, it possesses four

Table 1 Genome features

General features			
Genome size (bp)	6,264,403		
G+C content	66.6%		
Coding regions	89.4%		
Stable RNA	0.4%		
RNA			
rRNA	16S	23S	5S
<i>rrnA</i>	722,096–723,631	724,103–726,993	727,136–727,255
<i>rrnB</i>	4,793,731–4,792,196	4,791,724–4,788,836	4,788,693–4,788,574
<i>rrnC</i>	5,269,259–5,267,724	5,267,252–5,264,362	5,264,219–5,264,100
<i>rrnD</i>	6,044,743–6,043,208	6,042,736–6,039,846	6,039,703–6,039,584
Transfer RNA*	63 species		
Non-classical RNA	4 species		
Coding sequences			
Confidence level†	ORFs (%)	Definition	
1	372 (6.7)	<i>P. aeruginosa</i> genes with demonstrated function	
2	1059 (19.0)	Strong homologues of genes with demonstrated function from other organisms	
3	1590 (28.5)	Genes with proposed function based on motif searches or limited homology	
4	769 (13.8)	Homologues of reported genes of unknown function	
4	1780 (32.0)	No homology to any reported sequences	
Total	5570 (100)		

* Transfer RNAs were identified with tRNAscan-SE⁴⁰.

† Each annotation includes a numerical confidence level indicating the basis of determination of the protein name. A complete list of predicted genes and their annotations is available at <http://www.pseudomonas.com>.

dicarboxylate permeases of the TRAP-T type (*E. coli* has only one), and has only two phosphotransferase system (PTS) sugar transporters—for fructose and *N*-acetylglucosamine (*E. coli* has more than twenty)¹⁷. Also, *P. aeruginosa* has no predicted sugar transporters of the major facilitator superfamily (MFS), although *E. coli* has more than twenty. The apparent lack of sugar transporters in *P. aeruginosa* correlates with the absence of an intact glycolytic pathway and with its aerobic, oxidative metabolism¹⁸.

β-Oxidative metabolism. In contrast to its limited ability to grow on sugars, *P. aeruginosa* can use a wide variety of other carbon compounds, and its genome provides insight into the molecular basis of this metabolic versatility. In addition to known oxidative enzymes and pathways, we found a substantial number of other genes encoding putative enzymes characteristic of β-oxidation, such as acyl-CoA dehydrogenase (25 genes) and enoyl-CoA hydratase/isomerase (16 genes). In contrast, *E. coli* contains four genes for acyl-CoA dehydrogenase and seven for enoyl-CoA hydratase/isomerase. With the exception of *M. tuberculosis*, no other sequenced genome contains such large numbers of these enzymes. The β-oxidative genes are often clustered with other genes encoding proteins that may have related functions, such as probable acyl-CoA thiolases, short-chain dehydrogenases, flavin-containing monooxygenases, or other oxidoreductases. In several cases, these gene clusters also contained genes for MFS transport proteins and outer membrane porins of the OprD family (see Supplementary Information).

Intrinsic drug resistance and efflux systems. *P. aeruginosa* is noted for its intrinsic resistance to many front-line antibiotics, due mainly to its low outer membrane permeability and to active efflux of antibiotics¹⁹. Four *P. aeruginosa* multidrug efflux systems have been reported, all of which are members of the resistance-nodulation-cell division (RND) family^{20,21}. We used BLASTP analysis to identify potential export systems in the PAO1 genome, and probable multidrug efflux systems were identified by a phylogenetic analysis of each family¹⁷. The *P. aeruginosa* genome appears to contain a large number of undescribed drug efflux systems, predominantly of the RND and MFS families (Fig. 3). The number of predicted drug

efflux systems from the MFS, small multi-drug resistance (SMR), ATP-binding cassette (ABC) and multidrug and toxic compound extrusion (MATE) families is similar to other organisms such as *E. coli*, *B. subtilis* and *M. tuberculosis*. However, *P. aeruginosa* contains many more predicted AcrB/Mex-type RND multidrug efflux systems (10 genes) than *E. coli* (4), *B. subtilis* (1) and *M. tuberculosis* (0). Each of the *P. aeruginosa* genes encoding a putative RND transport protein is adjacent to a gene for a probable membrane fusion protein; most RND loci also contain genes for outer membrane proteins of the OprM family (see Supplementary Information).

Protein secretion. *P. aeruginosa* secretes several virulence factors, including toxins, lipases and proteases. Four pathways of protein secretion have been described for Gram-negative bacteria²², and three of these were evident in *P. aeruginosa*. The prototypic type I system in *P. aeruginosa*, which directs secretion of alkaline protease (encoded by *aprA*), consists of the ABC transport protein *aprD*, the membrane fusion protein *aprE* and the OprM-family outer membrane protein *aprF*. The PAO1 genome appears to contain four additional type I systems. One of these clusters (PA3404–PA3408) is homologous to the haem acquisition system (Has) of *Serratia marcescens*. A sixth *aprF* homologue (PA4974) was not clustered with genes for other putative transport proteins. This gene was most similar in sequence to *tolC*, which encodes an *E. coli* outer membrane protein involved in haemolysin secretion.

A type II secretion system (general secretion pathway) is encoded by the *xcp* gene cluster (PA3095–PA3105) and the unlinked *pilD/xcpA* gene (PA4528)²³. An unexpected finding was that many of the *xcp* genes have homologues elsewhere on the chromosome, including four additional homologues of *xcpQ* and *xcpR*. Type III secretion systems are found in many plant and human pathogens and are responsible for contact-dependent delivery of proteins into the cytoplasm of host cells. *P. aeruginosa* contains a single type III secretion system (PA1690–PA1725) which secretes several proteins including exoenzymes S, T and Y²⁴.

Other potential surface molecules. The genome of *P. aeruginosa* PAO1 contains two extremely long open reading frames, PA2462 (5,628 amino acids) and PA41 (3,536 amino acids). Each of these ORFs has sequence similarity to proteins that are adhesins in other bacterial pathogens, the filamentous haemagglutinin (FhaB) of *Bordetella pertussis*, and the HMW1A/HMW2A adhesins

Table 2 Functional classes of predicted genes

Function class	ORFS	% of ORFS
Adaptation, protection (for example cold shock proteins)	60	1.1
Amino acid biosynthesis and metabolism	150	2.7
Antibiotic resistance and susceptibility	19	0.3
Biosynthesis of cofactors, prosthetic groups and carriers	119	2.1
Carbon compound catabolism	130	2.3
Cell division	26	0.5
Cell wall, LPS	83	1.5
Central intermediary metabolism	64	1.1
Chaperones & heat shock proteins	52	0.9
Chemotaxis	43	0.8
DNA replication, recombination, modification and repair	81	1.5
Energy metabolism	166	3.0
Fatty acid and phospholipid metabolism	56	1.0
Membrane proteins	7	0.1
Motility & attachment	65	1.2
Nucleotide biosynthesis and metabolism	60	1.1
Protein secretion/export apparatus	83	1.5
Putative enzymes	409	7.3
Related to phage, transposon or plasmid	38	0.7
Secreted factors (toxins, enzymes, alginate)	58	1.0
Transcription, RNA processing and degradation	45	0.8
Transcriptional regulators	403	7.2
Translation, post-translational modification, degradation	149	2.7
Transport of small molecules	555	10.0
Two-component regulatory systems	118	2.1
Hypothetical	1,774	31.8
Unknown (conserved hypothetical)	757	13.6
Total	5,570	100.0

On the basis of known function or homology to genes of known function, PAO1 ORFs were assigned to one of 25 functional categories derived from those used for functional classification of *E. coli* ORFs¹¹. Where no functional inferences were possible, genes with homology to genes of unknown function predicted in other genomes were designated 'unknown', and genes without strong homology to any previously reported sequence were designated 'hypothetical'.

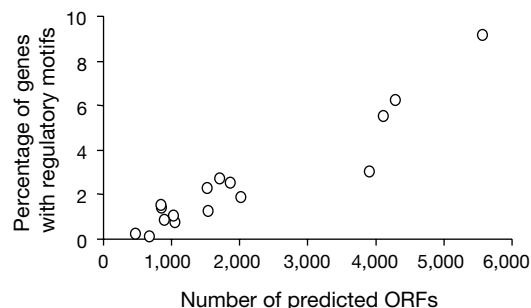


Figure 2 The percentage of genes with regulatory motifs increases with the size of the genome. Each sensor or regulatory family model was extracted from the Pfam 5.2 database¹⁵ and analysed against a database containing the combined predicted ORFs of each of the 22 genome sequences listed below. For each genome, the total number of ORFs identified with a probability of less than 10^{-4} as containing any of the regulatory motifs was divided by the number of predicted ORFs in that genome to calculate the percentage regulatory genes. The genomes analysed were *M. genitalium* (480 predicted ORFs), *M. pneumoniae* (677), *R. prowazekii* (834), *B. burgdorferi* (850), *C. trachomatis* (894), *T. pallidum* (1031), *C. pneumoniae* (1052), *A. aeolicus* (1522), *H. pylori* 26695 (1553), *H. influenzae* (1709), *M. jannaschii* (1715), *P. abyssi* (1765), *T. maritima* (1846), *M. thermoautotrophicum* (1869), *N. meningitidis* MC58 (2025), *P. horikoshii* (2064), *A. fulgidus* (2407), *Synechocystis* PCC6803 (3169), *M. tuberculosis* (3918), *B. subtilis* (4100), *E. coli* (4289), *P. aeruginosa* (5570).

of *Haemophilus influenzae*^{25,26}. PA2462 is adjacent to an ORF (PA2461) with strikingly abnormal codon usage and a G+C content of only 38.5%. Further, in addition to the genes known to be involved in synthesis of lipopolysaccharides, we noted three other genetic loci that may be involved in the synthesis of extracellular polysaccharides (see Supplementary Information). For example, a seven-gene cluster (PA1385–PA1391) adjacent to the *galE* gene includes genes for four probable glycosyl transferases and an ABC transport protein similar to putative carbohydrate-exporting proteins encoded in O-antigen biosynthetic gene clusters in other organisms. This seven-gene cluster is in a region with lower G+C content than the surrounding genes.

Chemosensing and chemotaxis. *P. aeruginosa* appears to have the most complex chemosensory systems of all the complete bacterial genomes, with four loci that encode probable chemotaxis signal-transduction pathways (see Supplementary Information). Of these, one (PA1456–PA1464) is similar in gene organization to the *Salmonella typhimurium* locus required for flagella-mediated swimming toward chemoattractants²⁷. A second (PA173–PA180) more closely resembles the gene organization seen in *Rhodobacter sphaeroides*²⁸. Each of the two remaining clusters has homology to the *che* genes from *E. coli* and to the *frz* genes from the non-flagellated gliding bacterium, *Myxococcus xanthus*²⁹. PA408–PA417 is required for twitching motility³⁰ and PA3702–PA3708 is as yet uncharacterized. *P. aeruginosa* undergoes chemotaxis toward a variety of sugars, amino acids, and inorganic phosphate, and away from thiocyanic and isothiocyanic esters^{31–34}. We identified 26 ORFs encoding probable chemotaxis sensory transducer proteins to mediate these responses.

Discussion

We propose that the large genome size and genetic complexity of *P. aeruginosa* reflect evolutionary adaptations permitting it to thrive in diverse ecological niches. Analysis of the complete genome sequence of *P. aeruginosa* reveals many clues regarding the basis of this versatility. *P. aeruginosa* has broad capabilities to transport, metabolize and grow on organic substances, numerous iron-side-phore uptake systems, and the enhanced ability to export

compounds (for example, enzymes and antibiotics) by a large number of protein secretion and RND efflux systems. *P. aeruginosa* potentially possesses four chemotaxis systems, at least one of which contributes to its ability to form biofilms³⁵. Thus this organism can readily move to more favourable conditions or consolidate and 'dig in' for persistent colonization of a particular microenvironment. Consistent with its increased genetic complexity, *P. aeruginosa* has the greatest percentage of genes devoted to command and control systems (for example, environmental sensors and transcriptional regulators) observed in a bacterial genome. These regulatory genes presumably modulate the diverse genetic and biochemical capabilities of this bacterium in changing environmental conditions.

P. aeruginosa infections are particularly difficult to treat because of intrinsic resistance to antibiotics. It would appear that, in the course of evolving the functional diversity required to compete with other microorganisms in a variety of environments, it developed mechanisms for resisting naturally occurring antimicrobial compounds. The efflux systems we identified could contribute to this intrinsic resistance. The effects of antimicrobials could also be mitigated by modulating expression of drug targets, enzymatic modifiers, transport systems and compensatory pathways. Indeed, the unusually large regulatory capability in *P. aeruginosa* may provide greater latitude for adaptive drug resistance through gene regulation than exists in other bacteria with smaller genomes. Furthermore, given its capacity to metabolize a wide variety of organic substrates, it is also possible that *P. aeruginosa* possesses greater potential for enzymatic modification and degradative drug resistance mechanisms than was thought. Therefore, the metabolic diversity, transport capabilities and regulatory adaptability that enable *P. aeruginosa* to thrive and compete with other microorganisms probably all contribute to its high intrinsic resistance to antibiotics. Knowledge of the complete genome sequence and encoded processes provides a wealth of information for the discovery and exploitation of new antibiotic targets, and hope for the development of more effective strategies to treat the life-threatening opportunistic infections caused by *P. aeruginosa* in humans. □

Methods

Sequencing and assembly

Strain PAO1, a wound isolate³⁶, was chosen as a strain prototype for sequencing because it is the most widely used *P. aeruginosa* laboratory strain and because physical and genetic maps were available^{12,13}. Strain PAO1 was obtained from B. Holloway's collection maintained in the laboratory of P. Phibbs, Univ. of Georgia. Our approach to the *P. aeruginosa* sequencing relied on standard data-collection and sequence-assembly methods. Most of the data comprised individual sequencing traces acquired from randomly picked M13 templates. The data were processed with the phred/phrap/consed package of base-calling, sequence-assembly, and finishing/editing software (<http://bozeman.mbt.washington.edu>)^{37,38}. At frequent intervals throughout the project, *de novo* phrap assemblies were carried out using all available data. The main requirement for achieving practical computation times was adequate real memory: on the full data set, the assembly required 4 h on a workstation with 4 gigabytes of memory. We employed dye-primer and dye-terminator chemistry in a 51:49 ratio to acquire 94,847 usable shotgun-sequencing traces. The shotgun traces provided 6.9-fold coverage of the genome in 'high-quality' base calls (that is, those with phred scores > 20, which corresponds to an error rate < 1%). The final raw-data set included an additional 1,604 'finishing' traces, most of which were obtained by priming with custom primers on M13 templates selected from the random-template collection. The purpose of most 'finishing' reads was to improve data quality in regions where the phred/phrap/consed software recognized that the consensus sequence had inadequate support. We also acquired 1,672 cosmid-end sequences from a collection of 836 cosmids that contained 40-kb inserts. The inserts in these cosmids covered 97% of the genome; hence, their end sequences provided a strong check on the validity of the final assembly.

Only 13 sites in the genome required specialized finishing procedures: four of these sites are the rDNA loci, which contain nearly exact copies of a 5.9-kb repeat, six are nearly identical copies of a 1.4-kb insertion sequence, and the other three also involve repeated sequences. The sequences at all 13 of these sites were obtained by sequencing PCR products that spanned the individual repeats. In addition to validating the assembly with cosmid-end sequences, we also monitored the base-pair accuracy of the final sequence in a variety of ways. One test involved full sequencing, by conventional methods, of two cosmids that contained widely spaced segments of the *P. aeruginosa* genome; these cosmids, which were selected at the beginning of the project, were brought to the highest

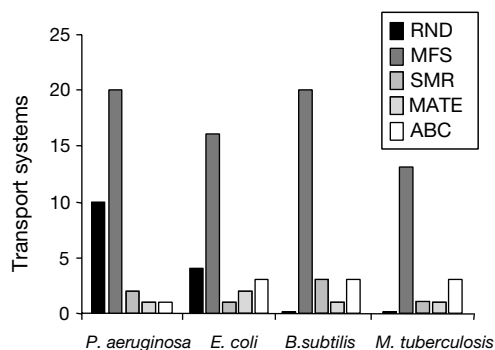


Figure 3 Comparison of the number of predicted drug efflux systems in *P. aeruginosa*, *E. coli*, *B. subtilis* and *M. tuberculosis*. For the last three organisms, these numbers are based on predictions taken from <http://www-biology.ucsd.edu/~ipaulesen/transport/>. Five types of multidrug efflux systems are analysed: resistance/nodulation/cell division family (RND; for example, *E. coli* AcrB), major facilitator superfamily (MFS; for example, *B. subtilis* Bmr), small multidrug resistance family (SMR; for example, *E. coli* EmrE), multidrug and toxic compound extrusion family (MATE; for example, *Vibrio parahaemolyticus* NorM)⁴¹ and ATP-binding cassette family (ABC; for example, *Lactococcus lactis* LmrA)¹⁷. Only family members that clearly clustered with known multidrug efflux systems were counted as probable multidrug efflux systems. For example, the number of RND multidrug efflux systems does not include members of this family that belong to the SecD/SecF protein excretion, the Czc metal efflux or the *M. tuberculosis* MmpL glycolipid efflux⁴² protein clusters, but only includes proteins belonging to the AcrB/Mex multidrug efflux protein cluster⁴³.

achievable quality standard by expert 'finishers'. In the final whole-genome assembly, which was entirely independent of the cosmid sequencing, we found no discrepancies with the 81,843 base pairs present in the two cosmids.

Gene predictions

Open reading frames were initially predicted by GeneMark.HMM (<http://genemark.biology.gatech.edu/GeneMark/whmm.cgi>)³⁹. Additional ORFs with homology to known genes were identified by BLASTX analysis. Predicted ORFs were reviewed individually by gene annotators for start-codon assignment based on additional contextual information such as the proximity of ribosomal binding sequence motifs and predicted signal peptides. □

Received 16 May; accepted 10 July 2000.

- Hardalo, C. & Edberg, S. C. *Pseudomonas aeruginosa*: assessment of risk from drinking water. *Crit. Rev. Microbiol.* **23**, 47–75 (1997).
- Costerton, J. W., Stewart, P. S. & Greenberg, E. P. Bacterial biofilms: a common cause of persistent infections. *Science* **284**, 1318–1322 (1999).
- Ahearn, D. G., Borazjani, R. N., Simmons, R. B. & Gabriel, M. M. Primary adhesion of *Pseudomonas aeruginosa* to inanimate surfaces including biomaterials. *Methods Enzymol.* **310**, 551–557 (1999).
- Bodey, G. P., Bolivar, R., Fainstein, V. & Jadeja, L. Infections caused by *Pseudomonas aeruginosa*. *Rev. Infect. Dis.* **5**, 279–313 (1983).
- Ball, C. A. *et al.* Integrating functional genomic information into the *Saccharomyces* genome database. *Nucleic Acids Res.* **28**, 77–80 (2000).
- Ewing, B. & Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genet.* **25**, 232–234 (2000).
- White, O. *et al.* Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* **286**, 1571–1577 (1999).
- Kunst, F. *et al.* The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249–256 (1997).
- Kaneko, T. *et al.* Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3**, 109–136 (1996).
- Cole, S. T. *et al.* Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome. *Nature* **393**, 537–544 (1998).
- Blattner, F. R. *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1474 (1997).
- Schmidt, K. D., Schmidt-Rose, T., Römmling, U. & Tümmler, B. Differential genome analysis of bacteria by genomic subtractive hybridization and pulsed field gel electrophoresis. *Electrophoresis* **19**, 509–514 (1998).
- Schmidt, K. D., Tümmler, B. & Römmling, U. Comparative genome mapping of *Pseudomonas aeruginosa* PAO with *P. aeruginosa* C, which belongs to a major clone in cystic fibrosis patients and aquatic habitats. *J. Bacteriol.* **178**, 85–93 (1996).
- Mahan, M. J. & Roth, J. R. Ability of a bacterial chromosome segment to invert is dictated by included maternal rather than flanking sequence. *Genetics* **129**, 1021–1032 (1991).
- Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **28**, 262–266 (2000).
- Hancock, R. E., Alm, R., Bina, J. & Trust, T. *Helicobacter pylori*: a surprisingly conserved bacterium. *Nature Biotechnol.* **16**, 216–217 (1998).
- Paulsen, I. T., Sliwinski, M. K. & Saier, M. H. Jr Microbial genome analyses: global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities. *J. Mol. Biol.* **277**, 573–592 (1998).
- Temple, L. M., Sage, A. E., Schweizer, H. P., & Phibbs, P. V. Jr in *Pseudomonas* (ed. Montie, T. C.) 35–72 (Plenum, New York, 1998).
- Hancock, R. E. Resistance mechanisms in *Pseudomonas aeruginosa* and other nonfermentative gram-negative bacteria. *Clin. Infect. Dis.* **27**, S93–S99 (1998).
- Westbrock-Wadman, S. *et al.* Characterization of a *Pseudomonas aeruginosa* efflux pump contributing to aminoglycoside impermeability. *Antimicrob. Agents Chemother.* **43**, 2975–2983 (1999).
- Nikaido, H. Antibiotic resistance caused by gram-negative multidrug efflux pumps. *Clin. Infect. Dis.* **27**, S32–S41 (1998).
- China, B. & Goffaux, F. Secretion of virulence factors by *Escherichia coli*. *Vet. Res.* **30**, 181–202 (1999).
- Bleves, S., Gerard-Vincent, M., Lazdunski, A. & Filloux, A. Structure-function analysis of XcpP, a component involved in general secretory pathway-dependent protein secretion in *Pseudomonas aeruginosa*. *J. Bacteriol.* **181**, 4012–4019 (1999).
- Yahr, T. L., Vallis, A. J., Hancock, M. K., Barbieri, J. T. & Frank, D. W. ExoY, an adenylate cyclase secreted by the *Pseudomonas aeruginosa* type III system. *Proc. Natl Acad. Sci. USA* **95**, 13899–13904 (1998).
- Domenighini, M. *et al.* Genetic characterization of *Bordetella pertussis* filamentous haemagglutinin: a protein processed from an unusually large precursor. *Mol. Microbiol.* **4**, 787–800 (1990).
- Noel, G. J., Barenkamp, S. J., St. Geme, J. W. III, Haining, W. N. & Mosser, D. M. High-molecular-weight surface-exposed proteins of *Haemophilus influenzae* mediate binding to macrophages. *J. Infect. Dis.* **169**, 425–429 (1994).
- Stock, J. B. & Surette, M. G. Chemotaxis. in *Escherichia coli and Salmonella: Cellular and Molecular Biology*. (ed. Neidhardt, F. C.) 1103–1129 (ASM, Washington, DC, 1996).
- Armitage, J. P. & Schmitt, R. Bacterial chemotaxis: *Rhodospirillum rubrum* and *Sinorhizobium meliloti*—variations on a theme? *Microbiology* **143**, 3671–82 (1997).
- McBride, M. J., Weinberg, R. A. & Zusman, D. R. "Frizzy" aggregation genes of the gliding bacterium *Myxococcus xanthus* show sequence similarities to the chemotaxis genes of enteric bacteria. *Proc. Natl Acad. Sci. USA* **86**, 424–428 (1989).
- Darzens, A. Characterization of a *Pseudomonas aeruginosa* gene cluster involved in pilus biosynthesis and twitching motility: sequence similarity to the chemotaxis proteins of enterics and the gliding bacterium *Myxococcus xanthus*. *Mol. Microbiol.* **11**, 137–153 (1994).
- Ohga, T., Masduki, A., Kato, J. & Ohtake, H. Chemotaxis away from thiocyanic and isothiocyanic esters in *Pseudomonas aeruginosa*. *FEMS Microbiol. Lett.* **113**, 63–66 (1993).
- Nelson, J. W. *et al.* Mucinophilic and chemotactic properties of *Pseudomonas aeruginosa* in relation to pulmonary colonization in cystic fibrosis. *Infect. Immun.* **58**, 1489–1495 (1990).
- Kato, J., Sakai, Y., Nikata, T. & Ohtake, H. Cloning and characterization of a *Pseudomonas aeruginosa* gene involved in the negative regulation of phosphate taxis. *J. Bacteriol.* **176**, 5874–5877 (1994).
- Moulton, R. C. & Montie, R. C. Chemotaxis by *Pseudomonas aeruginosa*. *J. Bacteriol.* **137**, 274–280 (1979).
- O'Toole, G. A. & Kolter, R. Flagellar and twitching motility are necessary for *Pseudomonas aeruginosa* biofilm development. *Mol. Microbiol.* **30**, 295–304 (1998).
- Holloway, B. W. Genetic recombination in *Pseudomonas aeruginosa*. *J. Gen. Microbiol.* **13**, 572–581 (1955).
- Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res.* **8**, 195–202 (1998).
- Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
- Lukashin, A. V. & Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26**, 1107–1115 (1998).
- Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
- Brown, M. H., Paulsen, I. T. & Skurray, R. A. The multidrug efflux protein NorM is a prototype of a new family of transporters. *Mol. Microbiol.* **31**, 394–395 (1999).
- Cox, J. S., Chen, B., McNeil, M. & Jacobs, W. R. Jr. Complex lipid determines tissue-specific replication of *Mycobacterium tuberculosis* in mice. *Nature* **402**, 79–83 (1999).
- Tseng, T.-T. *et al.* The RND permease superfamily: an ancient, ubiquitous and diverse family that includes human disease and development proteins. *J. Molec. Microbiol. Biotech.* **1**, 107–126 (1999).
- Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

X. Q. Pham managed the later stages of the shotgun, closure and finishing phases of the genome sequencing; A. Erwin did essential work in the annotation of the sequence and preparation of this manuscript; S. Mizoguchi managed the genome informatics databases, web site and analysis tools development; Kim Smith managed much of the primary-shotgun data collection; and F. Brinkman managed the *P. aeruginosa* collaborative annotation project. A total of 1741 genes were annotated by 61 expert volunteers from the *Pseudomonas* research community (L. Adewoye, A. M. Antonio, S. K. Arora, M. Bailey, S. Beatson, W. Bitter, R. Blakeley, C. O. Carranza, P. Cooke, P. Cornelis, L. Croft, T. de Kievit, R. De Mot, B. Erni, M. Eschbach, G. Fichant, C. J. Fields, A. Filloux, J. Campos-Garcia, P. Hager, T. Hayashi, R. Herzog, B. Huang, V. Huang, B. H. Iglewski, Y. Itoh, K.-E. Jaeger, D. Jahn, Z. Jalil, J. Kato, M. Kertesz, A. M. Kropinski, J. Lam, I. Lamont, A. Lazdunski, P. C. Lau, R. Levesque, K. Mathe, J. Mattick, J. Nezezon, M. Ochs, G. O'Donovan, L. Passador, P. Phibbs, Y. Quentin, A. Rodrigue, P. H. Roy, M. Saier, F. Sanschagrin, M. J. Schurr, P. C. Seed, T. H. M. Smits, G. Soberón-Chávez, L. G. Treviño, B. Tümmeler, J. van Beilen, M. Vasil, C. Whitchurch, K. Williams, K. Wong and B. Worobec). We thank P. Green and B. Ewing for assistance with the whole-genome phrap assemblies. Major funding for this project was provided by the Cystic Fibrosis Foundation and Pathogenesis Corporation. Funding for the organization of the PseudoCAP project was from the Canadian, French and German Cystic Fibrosis Foundations. R.E.W.H. was recipient of the Medical Research Council of Canada (MRC) Distinguished Scientist Award and received funding from the MRC.

Correspondence and requests for materials should be addressed to M.V.O. e-mail: mvo@u.washington.edu). The complete nucleotide sequence and sequences of predicted ORFs have been deposited with GenBank (accession no. AE004091). Sequence data, annotations and details of many of the analyses described here are also available on the *P. aeruginosa* genome web site, <http://www.pseudomonas.com>.

Resonance as a measure of pairing correlations in the high- T_c superconductor $\text{YBa}_2\text{Cu}_3\text{O}_{6.6}$

Pengcheng Dai*, H. A. Mook*, G. Aeppli†, S. M. Hayden‡ & F. Doğan§

* Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6393, USA

† NEC Research Institute, Princeton, New Jersey 08540, USA

‡ H. H. Wills Physics Laboratory, University of Bristol, Bristol BS8 1TL, UK

§ Department of Materials Science and Engineering, University of Washington, Seattle, Washington 98195, USA

One of the most striking properties of the high-transition-temperature (high- T_c) superconductors is that they are all derived from insulating antiferromagnetic parent compounds. The intimate relationship between magnetism and superconductivity in these copper oxide materials has intrigued researchers from the outset^{1–4}, because it does not exist in conventional superconductors. Evidence for this link comes from neutron-scattering experiments that show the unambiguous presence of short-range antiferromagnetic correlations (excitations) in the high- T_c superconductors. Even so, the role of such excitations in the pairing mechanism for superconductivity is still a subject of controversy⁵. For $\text{YBa}_2\text{Cu}_3\text{O}_{6+x}$ where x controls the hole-doping level, the most prominent feature in the magnetic excitation spectrum is a sharp resonance (refs 6–11). Here we show that for underdoped $\text{YBa}_2\text{Cu}_3\text{O}_{6.6}$, where x and T_c are below their optimal values, modest magnetic fields suppress the resonance significantly, much more so for fields approximately perpendicular to the CuO_2 planes than for parallel fields. Our results indicate that the resonance measures pairing and phase coherence, suggesting that magnetism plays an important role in high- T_c superconductivity. The persistence of a field effect above T_c favours mechanisms in which the superconducting electron pairs are pre-formed in the normal state of underdoped copper oxide superconductors^{12–14}, awaiting transition to the superconducting state.

Neutrons carry magnetic dipoles that are scattered by the electron spins in the sample with a probability—dependent on the neutron direction and energy—directly proportional to the Fourier transform in space and time of the two-spin correlation function, $S(\mathbf{Q}, \omega)$. The wavevector $\mathbf{Q}$ and energy $\hbar\omega$ are simply the difference between the momenta and energies, respectively, of the incident and scattered neutrons. $\text{YBa}_2\text{Cu}_3\text{O}_{6+x}$ (denoted as $(123)\text{O}_{6+x}$) with x of 0.6 was selected because it is a thoroughly characterized high- T_c superconductor that can be fabricated as the large single crystals required for neutron measurements. Equally important, thermodynamic data indicate that even modest fields have a large effect on the electronic entropy in the superconducting phase^{15,16}. Thus, if the electronic entropy changes around T_c are mostly due to spin excitations¹¹, they must respond to an external magnetic field in qualitatively the same way¹⁷.

The compound $(123)\text{O}_{6+x}$ contains pairs of CuO_2 layers separated by charge reservoir layers. We label wavevectors $\mathbf{Q} = (q_x, q_y, q_z)$ in units of $\text{\AA}^{-1}$ as $(H, K, L) = (q_x a/2\pi, q_y b/2\pi, q_z c/2\pi)$ in the reciprocal lattice units (r.l.u.) appropriate for the orthorhombic unit cell, for which the lattice parameters (in $\text{\AA}$) are $a = 3.83$, $b = 3.88$ ($\approx a$) and $c = 11.743$ if $x = 0.6$. For $x = 0$, $(123)\text{O}_{6+x}$ is an antiferromagnetic insulator where the spin on each Cu^{2+} ion is antiparallel to its nearest neighbours, both in its layer and in the immediately adjacent layer¹⁸. As the oxygen content x is increased, the material becomes a metal and eventually, near $x = 1$, an excellent high-temperature ($T_c \approx 93$ K) superconductor. At x near 0.6, $(123)\text{O}_{6+x}$ is also a good superconductor with T_c reduced to about 60 K. During the evolution from insulator to superconductor, static antiferromagnetism is

lost while magnetic (spin) fluctuations centred at the $x = 0$ antiferromagnetic Bragg positions—often referred to^{6–11} as $\mathbf{Q}_0 = (\pi, \pi)$ —persist. For highly doped $(123)\text{O}_{6+x}$, the most prominent feature in the spin fluctuation spectrum is a sharp resonance that appears below T_c at an energy of 41 meV (refs 6–8). When scanned at fixed frequency as a function of wavevector, the sharp peak is centred at (π, π) (refs 6–8) and its intensity is unaffected by a 11.5-T field in the ab -plane¹⁹. In our underdoped $(123)\text{O}_{6.6}$ ($T_c = 62.7$ K)⁹, the resonance occurs at 34 meV and is superposed on a continuum which is gapped at low energies¹¹. For frequencies below

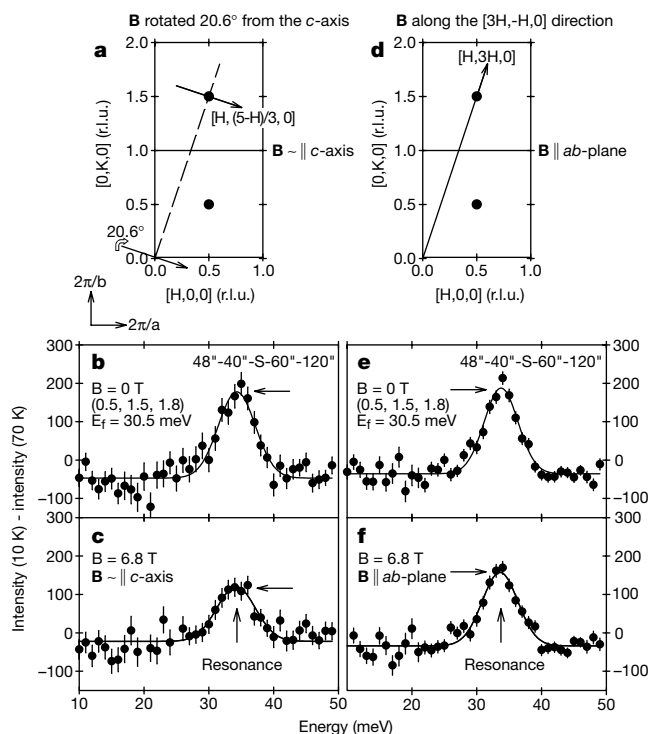


Figure 1 Reciprocal lattice diagram and neutron scattering results. For the experiment, we used the HB-3 triple-axis spectrometer at the High-Flux Isotope Reactor of Oak Ridge National Laboratory. The collimations were, proceeding from the reactor to the detector, $48^\circ\text{--}40'\text{--}60'\text{--}120'$ (full-width at half-maximum), and the final neutron energy was fixed at either $E_i = 30.5$ meV or 14.78 meV. The monochromator, analyser and filters were all pyrolytic graphite. The low-temperature data were collected after field cooling the sample from 200 K. **a**, To determine the effect of a field along the c -axis on the acoustic spin fluctuations, we first align the crystal in the $[H, K, 0]$ zone. The sample is then rotated α degrees around the $[3H, -H, 0]$ direction, where $\alpha = \arctan[(2\pi L/c)/(2\pi\sqrt{(H^2 + (3H)^2)/a})] \approx 20.6^\circ$ for $H = 0.5$ and $L = 1.8$ r.l.u., such that the horizontal scattering plane is spanned by wavevectors along the $[1, 3, \tan\alpha\sqrt{10}c/a]$ and $[3, -1, 0]$ directions. In this geometry, we can perform both constant- Q (panels **b** and **c**) and constant-energy (Fig. 2a–h) scans at the acoustic-mode intensity maximum ($L = 1.8$ r.l.u.) while still keeping 94% of the applied (vertical) field along the c -axis ($\mathbf{B} \sim \parallel c$ -axis). To see the influence of a field along c on the intensity of the optical mode, we oriented the c -axis of the crystal perpendicular to the scattering plane ($\mathbf{B} \parallel c$ -axis). This geometry allows any wavevector of form $(H, K, 0)$, that is, at the $L = 0$ intensity maximum of the optical mode, to be reached. **d**, Conventional $[H, 3H, L]$ zone geometry^{8–10} where the applied field is in the ab -plane along the $[3, -1, 0]$ direction. **e**, **f**, Difference spectra of the neutron intensities between $T = 10$ K ($< T_c$) and $T = 70$ K ($T_c + 7.3$ K) at wavevector $\mathbf{Q} = (0.5, 1.5, 1.8)$ for zero and 6.8-T field in the $\mathbf{B} \sim \parallel c$ -axis geometry. **e**, **f**, Similar data in the $\mathbf{B} \parallel ab$ -plane geometry. Because of the differences in sample geometries, the actual intensity gain of the resonance in the $\mathbf{B} \parallel ab$ -plane experiment is slightly higher than that of the $\mathbf{B} \sim \parallel c$ -axis case at zero field for the same incident beam current as measured by monitor (300 monitor counts correspond to ~ 8 min per point at $\hbar\omega = 34$ meV). We normalized the intensity gain of the resonance in these two experiments at zero field, and multiplied the finite field case by the same scale factor. Data in Figs 1–4 were obtained from $\text{YBa}_2\text{Cu}_3\text{O}_{6.6}$.

the resonance, the continuum is actually strongest at a quartet of incommensurate positions disposed symmetrically about (π, π) (refs 20, 21). Both the resonance and the incommensurate fluctuations appear in the acoustic channel for the bilayers, meaning that they are due to spin correlations that are antiferromagnetic between the planes. The corresponding neutron-scattering intensities exhibit a modulation $\sin^2(\pi dL/c)$, where d (3.342 Å) is the spacing between the nearest-neighbour CuO_2 planes, along c . In contrast, optical fluctuations are associated with ferromagnetic correlations between planes, and vary as $\cos^2(\pi dL/c)$ along c . The optical fluctuations for $x = 0.6$ have a gap of about 50 meV, well above the acoustic gap of 20 meV (ref. 11).

We mounted the crystal in three different orientations inside a split coil (to provide neutron beam access) 7-T vertical-field superconducting magnet to establish the effects of a field that was orientated in three directions. These directions were approximately along the c -axis ($\mathbf{B} \parallel c$ -axis, Fig. 1a; to investigate acoustic fluctuations), along the c -axis ($\mathbf{B} \parallel c$ -axis; to investigate optical fluctuations), and in the ab -plane ($\mathbf{B} \parallel ab$ -plane, Fig. 1d; to investigate acoustic fluctuations). Raw neutron-scattering data for $(123)\text{O}_{6+x}$ over the large frequency range covered here contain considerable non-magnetic background. The background, however, is weakly temperature-dependent for $T < 70$ K, which makes temperature difference spectra a useful guide to the magnetic contributions. Figure 1b shows the effect of superconductivity on $S(Q, \omega)$ at $Q = (1/2, 3/2, 1.8)$ in the form of difference spectra between 10 K ($< T_c$) and 70 K ($T_c + 7.3$ K). At zero field ($B = 0$ T), we reproduce previous results⁹ in both the $\mathbf{B} \parallel c$ -axis (Fig. 1b) and the $\mathbf{B} \parallel ab$ -plane (Fig. 1e) sample orientations. The data reveal the positive peak at 34 meV due to the increased resonance amplitude below T_c . Imposition of a 6.8-T field nearly along the c -axis results in the qualitatively similar temperature difference spectrum (Fig. 1c). However, the intensity gain of the

resonance is suppressed by $\sim 30\%$ compared to that at zero field (Fig. 1b). This is remarkable because the magnitude (6.8 T) of the applied field is much less than the upper critical field ($B_{c2} \approx 45$ T) for the material¹⁶, and the corresponding Zeeman (magnetic) energy is, at $\pm g\mu_B B$ (≈ 0.8 meV, assuming the Lande factor $g = 2$ and the resonance measures singlet-to-triplet transition with triplet spin $s = 1$), much smaller than the 34-meV resonance energy and the thermal energy $k_B T_c$ (≈ 6 meV).

In common with other layered superconductors, the thermodynamic properties of the copper oxides are strongly anisotropic with respect to the direction of the applied magnetic field¹⁶. A field perpendicular to the conducting planes generally causes a stronger suppression of the superconductivity than one applied within the plane. To see whether the resonance intensity depends similarly on field direction, we imposed a 6.8-T field parallel to the ab -plane (Fig. 1f). A noticeable, but much smaller, suppression ($\leq 10\%$) of the resonance is observed. Thus, the anisotropy of the field effect on the resonance reflects the anisotropy of other physical superconducting properties. This is the clearest evidence to date that the resonance is very sensitive to the superconducting pairing.

Although the constant- Q difference spectra (Fig. 1) are excellent for determining the intensity gain of the resonance from the normal to the superconducting state, such measurements do not provide information about how the field affects the momentum distribution of the excitations. We therefore scanned the wavevector in the $\mathbf{B} \parallel c$ -axis geometry (Fig. 1a) at energy transfers below, at, and above the resonance. At 150 K, the constant-energy scans at the resonance energy (Fig. 2a) show a broad peak centred at (π, π) on a sloped linear background. There are no observable intensity differences in the profiles for $B = 0$ and 6.8 T. On cooling to 70 K (Fig. 2b), the

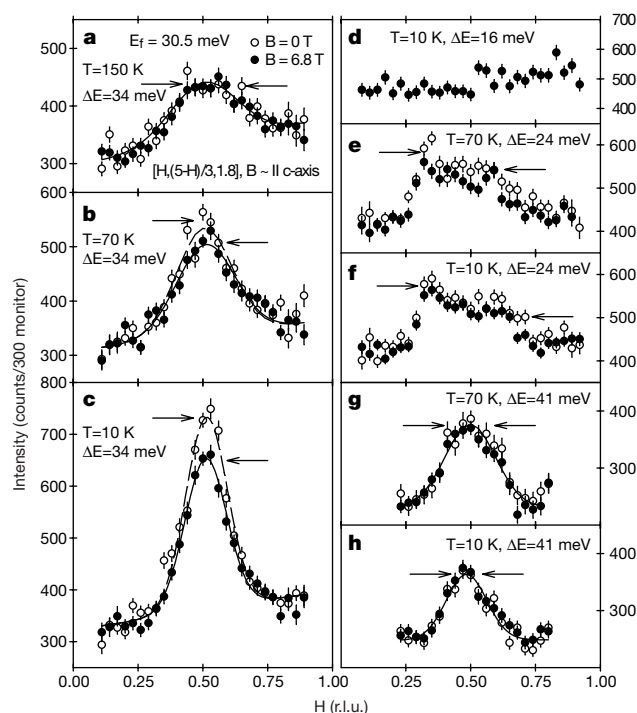


Figure 2 Effect of a magnetic field on the momentum dependence of the spin excitations. The figure shows constant-energy scans of the neutron scattering intensity as a function of wavevector in the $\mathbf{B} \parallel c$ -axis geometry. **a–c**, Results at the excitation energy of 34 meV; **d**, 16 meV; **e, f**, 24 meV; and **g, h**, 41 meV. The open and filled circles represent data for identical scans at zero and 6.8-T field, respectively. The dashed and solid lines are gaussian fits to the data.

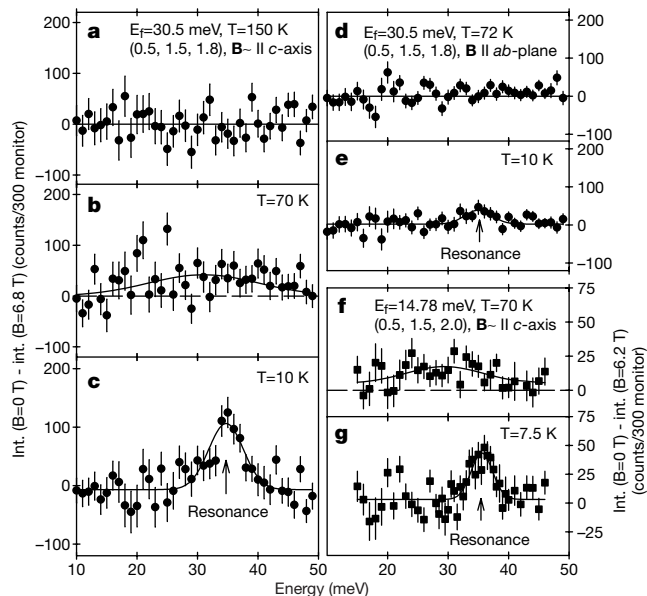


Figure 3 Effect of a magnetic field on the energy dependence of the spin excitations. The figure shows difference spectra of the neutron scattering intensities between zero and finite magnetic field. Panels **a–c** and **d–e** show the low-resolution results obtained with $E_i = 30.5$ meV at wavevector $Q = (0.5, 1.5, 1.8)$ in the $\mathbf{B} \parallel c$ -axis and $\mathbf{B} \parallel ab$ -plane geometries, respectively. High-resolution measurements in the $\mathbf{B} \parallel c$ -axis geometry with $E_i = 14.78$ meV are shown in **f** and **g**. The scattering should be centred around zero as shown by the solid lines in **a** and **d** if spin fluctuations are not affected by the field. The solid lines in **c** and **e–g** are unconstrained least-squares fits to the data by gaussians on floating constant backgrounds. The solid line in **b** is a guide to the eye. Positive scattering (above zero) indicates a net loss in the acoustic magnetic spectral weight when the field is applied. Strong phonon scattering at about 20 meV in the raw constant- Q scans yielded the large error bars in the difference spectra around that energy.

resonance narrows in width and grows in intensity, but is also clearly suppressed at 6.8 T. At 10 K (Fig. 2c), the suppression of the resonance at 6.8 T is more severe. Direct inspection and fits of gaussian profiles to the data (Fig. 2a–c) indicate no appreciable field-induced change in the momentum dependence of the resonance peak. In contrast, the thermal fluctuations associated with warming from 10 to 70 K yield a decrease in Q -integrated intensity that is numerically similar to that found on application of a 6.8-T field at 10 K, and makes the profile considerably broader.

Below the 20-meV spin gap in the superconducting state¹¹, the constant-energy scan at 16 meV in the 6.8-T field (Fig. 2d) is featureless, that is, there are no strong field-induced antiferromagnetic correlations at this energy. At 24 meV, where incommensurate spin fluctuations were observed^{20,21}, the field seems to suppress the constant-energy profiles in both the normal (Fig. 2e) and the superconducting state (Fig. 2f). The effect is more difficult to see than at the resonance energy, although it could represent a similar fractional change in a considerably smaller signal above background. Finally, when we move to an energy just above the resonance, we discern no field-induced change in the Q -dependent profiles at both 70 K and 10 K (Fig. 2g and h).

To demonstrate the effect of a magnetic field on the energy dependence of the excitations in $(123)\text{O}_{6.6}$, we show the experimentally measured difference spectra, $S(Q_0, \omega, B = 0 \text{ T}) - S(Q_0, \omega, B = 6.8 \text{ T})$, for $B \parallel c$ -axis (Fig. 3a–c) and in the ab -plane (Fig. 3d and e) at different temperatures. As we expect the lattice contribution (phonons) to the scattering to be field-insensitive, this procedure will yield the net change of the magnetic intensity induced by the field. At 150 K, a 6.8-T field has no effect on the spin fluctuations in the measured energy range for $B \parallel c$ -axis (Fig. 3a). On cooling to 70 K, the difference spectra show a broad peak for $B \parallel c$ -axis (Fig. 3b) and no discernible feature for $B \parallel ab$ -plane (Fig. 3d). Higher-resolution measurements in the $B \parallel c$ -axis geometry at 6.2-T field (Fig. 3f) confirm the result of Fig. 3b. In the superconducting state (Fig. 3c, e and g), the difference spectra show a sharp peak at 34 meV for both field orientations, again confirming that the resonance is the feature most obviously affected by the magnetic field for spin fluctuations below 50 meV. In addition, the sharpness here contrasts with the diffuse nature of the difference just above T_c , thus showing the sensitivity of the resonance width, or

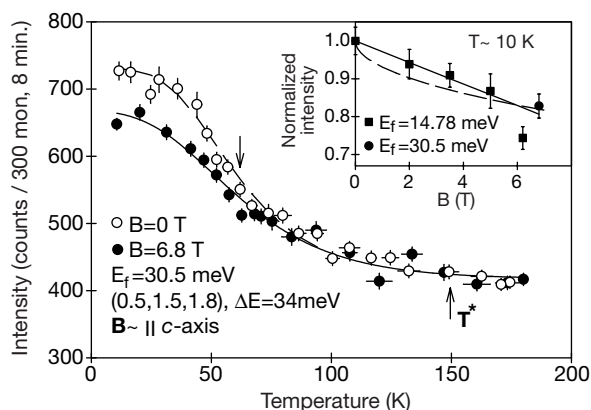


Figure 4 Effect of a magnetic field on the temperature dependence of the resonance and the field dependence of the resonance at about 10 K. The data are collected in the $B \parallel c$ -axis geometry. The open and filled circles show results at $Q = (0.5, 1.5, 1.8)$ and $\hbar\omega = 34$ meV for zero and 6.8-T field, respectively. The pseudogap temperature T^* and T_c of the sample are indicated by arrows. The solid and dashed lines are guides to the eye. The inset shows the field dependence of the normalized resonance intensity at approximately 10 K. The filled squares are high-resolution data at $Q = (0.5, 1.5, 2)$ with $E_f = 14.78$ meV. The solid line corresponds to $I/I_0 = 1 - (B/B_{\text{char}})$ with $B_{\text{char}} = 36$ T, while the dashed line represents $I/I_0 = 1 - (B/B_{\text{char}})^{1/2}$, where $B_{\text{char}} = 208$ T.

decay rate, to true bulk superconductivity even in a material with an extended pseudogap regime above T_c . Simple summation as well as least-squares fits to the data (Fig. 3f and g) indicate that the energy-integrated effect of the magnetic field just above T_c is comparable to that well below T_c . The narrowing of the widths in the field-induced difference spectra together with the approximate preservation of weight in these spectra suggest that the resonance width depends strongly on the phase coherence time, while its integrated weight measures the pair density.

In Fig. 4 we plot the temperature dependence of the neutron scattering intensity at the resonance energy at zero field and at a field of 6.8 T. Most of the growth in the field-dependent peak signal tracks the growth in the resonance itself on entering the superconducting state. The inset shows the field dependence of the normalized intensity. The solid line is a linear fit to the data—that is, $I/I_0 = 1 - B/B_{\text{char}}$ with $B_{\text{char}} = 36$ T. Because B_{char} is not far from the upper critical field $B_{c2} = 45$ T. Estimated for our sample, it is reasonable to associate the sharp resonance peak with the superconducting volume fraction, which for simple superconductors is proportional to $1 - B/B_{c2}$.

Figures 2 and 3 indicate that wherever we look at fixed temperature, the external field either has no effect or reduces the magnetic scattering signal. The total moment sum rule^{22,23} states that the magnetic structure factor, when integrated over all wavevectors and energies, should be a temperature- and field-independent constant. Therefore, compensatory increases in signal should occur elsewhere in Q - ω space. Possible destinations for the lost spectral weight include elastic scattering, new or softened optic modes, a subgap continuum, and other continuum scattering spread sufficiently over Q and ω space so as to be unobservable at any particular Q and ω . Searches for the most obvious elastic scattering—that due to the proposed field-induced antiferromagnetic order in copper oxide superconductors²⁴—have been fruitless. Scans in the $[H, K, 0]$ plane with $B \parallel c$ -axis revealed no evidence for optic mode scattering at $\hbar\omega = 34, 41$ and 50 meV apart from that already present at zero field. In addition, we have not yet found field-induced subgap scattering (Fig. 2d). Although our searches for the missing spectral weight were by no means exhaustive, a 6.8-T field does not induce obvious new excitations around (π, π) with energies between 10 and 50 meV that are big enough to compensate the spectral weight loss of the resonance. We note that similar sum rule violations also exist for spectra collected using angle resolved photoemission²⁵, scanning tunnelling microscopy²⁶, and infrared spectroscopy²⁷ in some high- T_c superconductors. However, we point out that an explanation of the superficially similar effects in these measurements must take into consideration the fact that different experimental techniques probe quite different correlations.

We have discovered that a very modest field applied to $(123)\text{O}_{6.6}$ yields a very significant reduction in the 34-meV resonance. Therefore, the resonance can be reduced either by warming or applying a magnetic field, with the highest amplitude and narrowest profiles (both in energy and wavevector) occurring in the bulk superconducting state. More surprising is the persistence of the effect above T_c (Figs 2–4). Although there are no explicit microscopic predictions about the effects of magnetic fields on the resonance, our finding has important implications for understanding the mechanism of high- T_c superconductivity. In particular, the larger c -axis field effect presents another challenge to the interlayer tunnelling theory, according to which the resonance should be strongly affected by the in-plane field that disrupts the coherent Josephson coupling along the c -axis²⁸. On the other hand, our data are consistent with mechanisms where the dominant loss of entropy on entering the superconducting state is due to the growth of magnetic correlations in the CuO_2 planes^{29–31}. A modest field reduces the resonance just as it reduces the specific heat anomaly¹⁵ near T_c , so that the temperature derivative of the resonance intensity tracks the specific heat not only as the hole density is

varied¹¹, but also as the magnetic field is imposed for fixed x (ref. 17). The anisotropic field effect reflects the anisotropy of other physical properties, such as the magnetic penetration depth and coherence length. Thus, the persistence of the effect above T_c would naturally be due to superconducting fluctuations or incoherent pairing, which also display this anisotropy in the normal state. Such fluctuations exist in models postulating a Bose–Einstein condensation of pre-formed pairs at T_c in the underdoped copper oxides^{12–14}. □

Received 16 November 1999; accepted 27 June 2000.

- Anderson, P. W. The resonating valence bond state in La_2CuO_4 . *Science* **235**, 1196–1198 (1987).
- Emery, V. J. Theory of high- T_c superconductivity in oxides. *Phys. Rev. Lett.* **58**, 2794–2797 (1987).
- Hirsch, J. E. Antiferromagnetism, localization, and pairing in a two-dimensional model for CuO_2 . *Phys. Rev. Lett.* **59**, 228–231 (1987).
- Schulz, H. J. Superconductivity and antiferromagnetism in the two-dimensional Hubbard model: scaling theory. *Europhys. Lett.* **4**, 609–615 (1987).
- Scalapino, D. J. The cuprate pairing mechanism. *Science* **284**, 1282–1283 (1999).
- Rossat-Mignod, J. *et al.* Neutron scattering study of the $\text{YBa}_2\text{Cu}_3\text{O}_{6+x}$ system. *Physica C* **185–189**, 86–92 (1991).
- Mook, H. A., Yethiraj, M., Aeppli, G., Mason, T. E. & Armstrong, T. Polarized neutron determination of the magnetic excitations in $\text{YBa}_2\text{Cu}_3\text{O}_7$. *Phys. Rev. Lett.* **70**, 3490–3493 (1993).
- Fong, H. F. *et al.* Phonon and magnetic neutron scattering at 41 meV in $\text{YBa}_2\text{Cu}_3\text{O}_7$. *Phys. Rev. Lett.* **75**, 316–319 (1995).
- Dai, P., Yethiraj, M., Mook, H. A., Lindemer, T. B. & Doğan, F. Magnetic dynamics in underdoped $\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$: direct observations of a superconducting gap. *Phys. Rev. Lett.* **77**, 5425–5428 (1996).
- Fong, H. F., Keimer, B., Milius, D. L. & Aksay, I. A. Superconducting-induced anomalies in the spin excitation spectra of underdoped $\text{YBa}_2\text{Cu}_3\text{O}_{6+x}$. *Phys. Rev. Lett.* **78**, 713–716 (1997).
- Dai, P. *et al.* The magnetic excitation spectrum and thermodynamics of high- T_c superconductors. *Science* **284**, 1344–1347 (1999).
- Uemura, Y. J. Bose–Einstein to BCS crossover picture for high- T_c cuprates. *Physica C* **282**, 194–197 (1997).
- Emery, V. J. & Kivelson, S. A. Importance of phase fluctuations in superconductors with small superfluid density. *Nature* **374**, 434–437 (1995).
- Lee, P. A. & Wen, X. G. Unusual superconducting state of underdoped cuprates. *Phys. Rev. Lett.* **78**, 4111–4114 (1997).
- Junod, A., Erb, A. & Renner, Ch. Specific heat of high temperature superconductors in high fields at T_c : from BCS to the Bose–Einstein condensation. *Physica C* **317**, 333–344 (1999).
- Junod, A. in *Studies of High Temperature Superconductors Vol. 19* (ed. Narlikar, A. V.) 1 (Nova Science, New York, 1996).
- Janko, B. Thermodynamic constraints on the magnetic field dependence of the neutron resonance in cuprate superconductors. Preprint cond-mat/9912073 at (<http://xxx.lanl.gov>) (1999; cited 6 Dec. 1999).
- Tranquada, J. M. *et al.* Neutron-diffraction determination of antiferromagnetic structure of Cu ions in $\text{YBa}_2\text{Cu}_3\text{O}_{6+x}$ with $x = 0.0$ and 0.15 . *Phys. Rev. Lett.* **60**, 156–159 (1988).
- Bourges, P. *et al.* High magnetic field dependence of spin fluctuations in $\text{YBa}_2\text{Cu}_3\text{O}_7$. *Physica B* **234–236**, 830–831 (1997).
- Dai, P., Mook, H. A. & Doğan, F. Incommensurate magnetic fluctuations in $\text{YBa}_2\text{Cu}_3\text{O}_{6.6}$. *Phys. Rev. Lett.* **80**, 1738–1741 (1998).
- Mook, H. A. *et al.* Spin fluctuations in $\text{YBa}_2\text{Cu}_3\text{O}_{6.6}$. *Nature* **395**, 580–582 (1998).
- Hayden, S. M. *et al.* Absolute measurements of the high-frequency magnetic dynamics in high- T_c superconductors. *Physica B* **241–243**, 765–772 (1998).
- Aeppli, G. *et al.* The weights of various features in the magnetic spectra of cuprates. *Phys. Status Solidi B* **215**, 519–522 (1999).
- Arovas, D. P., Berlinsky, A. J., Kallin, C. & Zhang, S. C. Superconducting vortex with antiferromagnetic core. *Phys. Rev. Lett.* **79**, 2871–2874 (1997).
- Anderson, P. W. Two crucial experimental tests of the resonating valence bond–Luttinger liquid interlayer tunneling theory of high- T_c superconductivity. *Phys. Rev. B* **42**, 2624–2626 (1990).
- Renner, Ch., Revaz, B., Genoud, J.-Y., Kadowaki, K. & Fischer, Ø. Pseudogap precursor of the superconducting gap in under- and overdoped $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+x}$. *Phys. Rev. Lett.* **80**, 149–152 (1998).
- Basov, D. N. *et al.* Sum rules and interlayer conductivity of high- T_c cuprates. *Science* **283**, 49–52 (1999).
- Yin, L., Chakravarty, S. & Anderson, P. W. The neutron peak in the interlayer tunneling model of high temperature superconductors. *Phys. Rev. Lett.* **78**, 3559–3562 (1997).
- Scalapino, D. J. & White, S. R. Superconducting condensation energy and an antiferromagnetic exchange-based pairing mechanism. *Phys. Rev. B* **58**, 8222–8224 (1998).
- Demler, E. & Zhang, S. C. Quantitative test of a microscopic mechanism of high-temperature superconductivity. *Nature* **396**, 733–735 (1998).
- Chakravarty, S. & Kee, H. Y. Measuring condensate fraction in superconductors. *Phys. Rev. B* **61**, 14821–14824 (2000).

Acknowledgements

We thank B. Janko, A. Junod, A. Kapitulnik, M. V. Klein, Ch. Renner and D. J. Scalapino for discussions. The work at ORNL was supported by the US DOE.

Correspondence and requests for materials should be addressed to P.D. (e-mail: piq@ornl.gov).

Quantum correlation among photons from a single quantum dot at room temperature

P. Michler*, A. Imamoglu*, M. D. Mason†, P. J. Carson†, G. F. Strouse† & S. K. Buratto†

* Department of Electrical and Computer Engineering, and Department of Physics, University of California, Santa Barbara, California 93106, USA

† Department of Chemistry and Biochemistry, University of California, Santa Barbara, California 93106, USA

Maxwell's equations successfully describe the statistical properties^{1,2} of fluorescence from an ensemble of atoms or semiconductors in one or more dimensions. But quantization of the radiation field is required to explain the correlations of light generated by a single two-level quantum emitter, such as an atom, ion or single molecule^{3–6}. The observation of photon antibunching in resonance fluorescence from a single atom unequivocally demonstrated the non-classical nature of radiation³. Here we report the experimental observation of photon antibunching from an artificial system—a single cadmium selenide quantum dot at room temperature. Apart from providing direct evidence for a solid-state non-classical light source, this result proves that a single quantum dot acts like an artificial atom, with a discrete anharmonic spectrum. In contrast, we find the photon-emission events from a cluster of several dots to be uncorrelated.

The physics of photon antibunching is easy to understand: if a two-level atom emits a photon at time $\tau = 0$, it is impossible for it to emit another one immediately after, because it is necessarily in the ground state. The next photon can only be emitted after a waiting time, which, under weak excitation conditions, is determined by the spontaneous emission time. The result is that there is a dead-time between successive photon-emission events, and the generated photon stream, in the absence of other sources of fluctuation, is sub-poissonian¹. Photon antibunching is easily washed away with increasing number of atoms. Conversely, we can consider strong photon antibunching as evidence that the source of the radiation field is a single anharmonic (for example, two-level) quantum system: if the spectrum of the emitter is harmonic or there is more than one anharmonic emitter, then the detection of the first

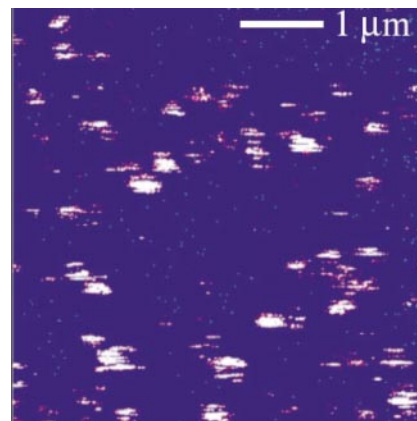


Figure 1 Photoluminescence image of CdSe/ZnS nanocrystals on a glass plate. Image was acquired by raster scanning the particle-covered glass plate through the laser focus ($\lambda = 488$ nm, FWHM was 300 nm) and collecting the photoluminescence onto a single-photon-counting avalanche photodiode. This 256×256 pixel image (4.9 ms per pixel) represents an $5 \times 5 \mu\text{m}^2$ field of view.

photon at $\tau = 0$ will no longer guarantee that the whole system is in its (radiatively inactive) ground state immediately after $\tau = 0$.

Photon correlation measurements on single quantum dots (QDs) were carried out by combining a Hanbury–Brown–Twiss photon coincidence set-up with a confocal scanning optical microscope (CSOM). All measurements were carried out at room temperature. To the best of our knowledge, this is the first QD experiment that goes beyond first-order optical coherence measurements¹.

Figure 1 shows a typical confocal image ($5 \times 5 \mu\text{m}$) of CdSe/ZnS nanoparticles on a glass plate. Each bright spot corresponds to the luminescence from either a single CdSe/ZnS quantum dot or a cluster of several dots. However, it is important to note that the smallest spot size ($\sim 300 \text{ nm}$) is defined by the resolution of the CSOM. As can be seen, bright spots with different sizes are visible and the smallest spots appear to blink 'on' and 'off' during the course of a scan⁷. This behaviour is thought to be the result of QD ionization, which will be discussed in more detail below.

The measured raw data photon correlation signals from a cluster of several QDs and a single CdSe/ZnS quantum dot are shown in Fig. 2. The single QD data shows a minimum of coincidence counts $n(\tau)$ around a pair separation time of $\tau = 0$, and an exponential increase of $n(\tau)$ for negative and positive values of τ . This is a clear signature of non-classical photon antibunching^{1,3}. In contrast, the photon correlation signal of the CdSe/ZnS cluster is flat. In this case many independently radiating QDs contribute to the signal.

Quantum dots are fairly complicated artificial atoms. Even for an isolated QD, the optical transitions have non-trivial signatures such as spectral diffusion and relatively large non-radiative broadening⁸. To model our photon correlation experiment, we use a simple three-state model for the QDs. We assume that the excited electron–hole pair relaxes incoherently before spontaneously emitting a photon: this relaxation process between the two excited states is much faster than other relevant rates⁹. In addition, we have introduced a fast, elastic dephasing rate for the coherences between the ground and excited states. Using the quantum regression theorem, we derive an expression for the normalized second-order correlation function¹ $g^{(2)}(\tau) = \langle : I(t)I(t+\tau) : \rangle / \langle I(t) \rangle^2$ (where $:$ indicates normal ordering and $I(t)$ is the measured intensity) in the weak-laser excitation regime

$$g^{(2)}(\tau) = 1 - e^{-(\Gamma + W_p)\tau}. \quad (1)$$

Here Γ is the electron–hole recombination rate and W_p is the effective pump rate into the radiatively active upper state. The fast dephasing of the emitting transition has no effect on $g^{(2)}(\tau)$ of a single QD. In contrast, for an ensemble of dots, the bosonic bunching effect decays with the dephasing time, which is estimated to be at least 10 ps using the narrowest observed linewidths⁸. This result indicates that with the time resolution of our experiment $\tau_{\text{res}} = 420 \text{ ps}$, we should expect to find uncorrelated photon counting events for an ensemble of dots (that is, $g^{(2)}(\tau) = 1$). The

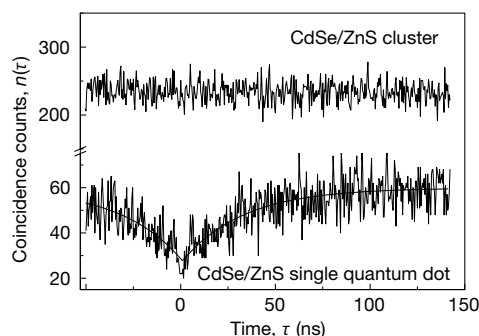


Figure 2 Measured distribution $n(\tau)$ of photon pair separation times τ for a CdSe/ZnS cluster and a single quantum dot. The line represents a fit to an exponential law and is described in the text.

experimental observations discussed earlier are fully consistent with this prediction. We remark that the measured $n(\tau)$ follows the autocorrelation $g^{(2)}(\tau)$ in the limit where the reciprocal of the average counting rate is much longer than the monitored time range. As this is always the case for our measurements, we calculate $g^{(2)}(\tau)$ by normalizing $n(\tau)$ to the long time average of 60 counts per channel.

The single QD data are fitted by $g^{(2)}(\tau) = 1 - ae^{-\tau/t_d}$ with decay time $t_d = 1/(\Gamma + W_p)$ and a being fit parameters. The factor a is introduced to take care of the stray-light background and residual biexcitonic effects. We obtain $t_d = 32 \pm 2 \text{ ns}$ which directly determines the single electron–hole pair lifetime under our experimental conditions ($\Gamma \gg W_p$). This value is of the same order of magnitude as independent lifetime measurements on ensembles of QDs. The fit also gives an estimate of $g^{(2)}(0) = 0.47 \pm 0.02$, which is a clear signature of non-classical photon antibunching from a single anharmonic quantum system.

The QD photoluminescence intensity versus time was recorded with an integration time of 2 ms per bin and typical scan durations of 5 to 30 min. Figure 3a and b shows the photoluminescence intensity time trace of a single CdSe/ZnS quantum dot in the long time range. An irregular and multi-level intensity emission is observed for the chosen size of the time bins. The background level of 3,000 counts s^{-1} is visible around 151 s in Fig. 3b where the CdSe quantum dot is in an 'off' state. This turning 'on' and 'off' behaviour (blinking) has been interpreted in terms of an Auger ionization model^{7,10}. In the case where two electron–hole pairs are excited in the QD, the recombination energy ($\sim 2.14 \text{ eV}$) of the first pair can be used to eject either the electron or the hole from the second pair into a deep trap of the surrounding matrix. A CdSe quantum dot ionized in this way is predominantly non-emissive, that is, in an 'off' state⁷. The bright state, that is, the 'on' state, is recreated either by recapturing the ejected carrier or through a second Auger process. The fast Auger rate is therefore not only responsible for the blinking of the QDs but also reduces the exciton-to-biexciton transition amplitude, making the virtual amplitude in the biexciton state small. However, one would still expect to find $g^{(2)}(0) > 0$ due to residual biexciton luminescence.

Figure 4 shows the calculated classical intensity autocorrelation function $G^{(2)}(\tau) = \langle I(t)I(t+\tau) \rangle / \langle I(t) \rangle^2$ for the time trace of Fig. 3. The signature of blinking at this timescale is the bunching behaviour ($G^{(2)}(0) > 1$). Clearly, the photon-bunching due to QD blinking can be described classically and has to be distinguished from the photon antibunching observed in the short time range (0–200 ns), which is due to the quantum nature of the emitter. The fact that the photon pair distribution remains flat for $100 \text{ ns} \leq \tau \leq 150 \text{ ns}$ is a good

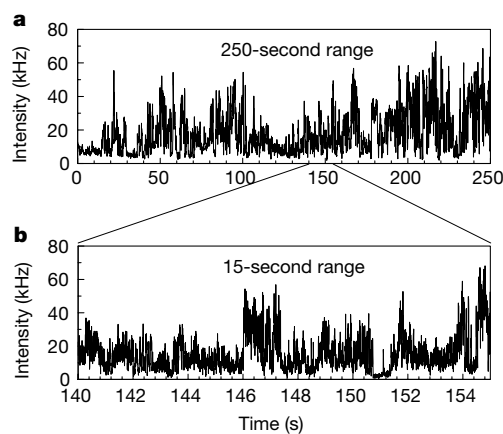


Figure 3 Time trace of photoluminescence intensity of a single CdSe/ZnS quantum dot on two different timescales. **a**, 250-second time range with 62.5 ms per bin. **b**, Expanded time range (15 seconds) with 3.9 ms per bin.

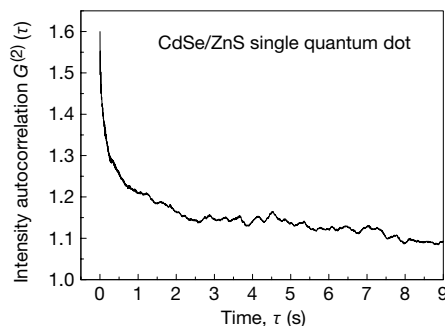


Figure 4 Intensity autocorrelation $G^{(2)}(\tau)$ obtained for the time trace shown in Fig. 3.

indication that blinking does not occur on the timescales of interest for photon antibunching.

Prior measurements on single QD photoluminescence and absorption have demonstrated the existence of discrete QD resonances⁸. However, just as the observation of discrete absorption lines in an atomic vapour cannot be taken as an evidence that the observed system consists of a single atom, these experiments, at least in principle, cannot rule out the existence of several QDs. In contrast, photon correlation measurements, such as the one reported here, provide a reliable method for deciding whether or not the observed system is a single anharmonic quantum emitter. In addition, owing to interactions with the lattice, a nanostructure that acts like an anharmonic emitter at cryogenic temperatures can be indistinguishable from a higher-dimensional system at room temperature. We have demonstrated that a CdSe quantum dot behaves as an anharmonic emitter even at room temperature. Our results constitute a first step in the study of quantum optical phenomena in semiconductor QDs. Demonstration of quantum dynamics in semiconductors at room temperature could lead to applications in quantum information processing and computation. □

Methods

Sample preparation

The CdSe/ZnS (core/shell) quantum dots were synthesized following high-temperature organometallic methods described in the literature^{11–13}. The resulting nanoparticles were capped with the organic ligand trioctylphosphine oxide (TOPO) and had a distribution with an average diameter of 4.1 nm and a r.m.s. (root mean square) width of 0.33 nm (8% size distribution). The single QD samples were prepared by spin-casting 30 μ l of a 0.5 nM solution of the QDs dissolved in hexanes onto a bare glass coverslip, which resulted in a mean separation of the QDs of approximately 1 μ m.

Experimental set-up

Optical pumping was performed using circularly polarized light of the 488-nm line of a continuous-wave Ar⁺ laser, generating electron–hole pairs in the excited states of the CdSe quantum dots. The exciting light was focused by a high numerical aperture (NA is 1.3) oil-immersion objective to a near-diffraction-limited (full-width at half-maximum (FWHM)) spot approximately 300 nm in diameter at the glass–sample interface. In order to minimize the generation of two electron–hole pairs simultaneously we used a low excitation intensity of about 250 W cm^{−2}. Typically, a QD was excited approximately every 1–10 μ s (ref. 14), whereas the photoluminescence decay time is of the order of 30 ns. Thus the probability of generating two electron–hole pairs was small. The photoluminescence from the QD was collected by the same objective and first passed through the excitation laser beam splitter and then through a holographic notch filter to block scattered laser light. The light was then split with a 50/50 non-polarizing beam splitter and the resulting two photon beams were focused onto the active areas of two single-photon-counting avalanche photodiodes (SPAD).

Measurement

Photoluminescence images of the nanocrystals were obtained by scanning the QD-covered glass plate through the laser focus and recording the number of counts with one of the SPADs. To measure the photon statistics of a selected QD (or a cluster of QDs), the glass plate was positioned where a single bright spot of typically 300 nm in diameter (resolution limited) was observed in the photoluminescence image. The number of pairs of photons $n(\tau)$ with arrival-time separations of τ was measured using the two SPADs for $\tau \leq \tau_{\max} = 200$ ns. The pulses from the two SPADs were used to start and stop a time-to-amplitude converter (TAC) where the time delay between the start and stop pulses (to within τ_{res}) was converted to a voltage amplitude. The SPADs exhibited the same counting

rate. An electronic delay (53 ns) was introduced in the stop channel in order to check the symmetry of the $n(\tau)$ signal and to avoid the effect of noise for small voltages in the TAC. The output pulses were fed into a multichannel analyser. To reduce the background contribution and therefore decrease the amount of uncorrelated light in the $n(\tau)$ measurement, the multichannel analyser was enabled only during the ‘on’ periods, that is, when the signal level was above a certain threshold.

Received 17 April; accepted 3 July 2000.

- Walls, D. F. & Milburn, G. J. *Quantum Optics* (Springer, Berlin, 1994).
- Hanbury-Brown, R. & Twiss, R. Q. Correlation between photons in two coherent beams of light. *Nature* **177**, 27–29 (1956).
- Kimble, H. J., Dagenais, M. & Mandel, L. Photon antibunching in resonance fluorescence. *Phys. Rev. Lett.* **39**, 691–694 (1977).
- Diedrich, F. & Walther, H. Nonclassical radiation of a single stored ion. *Phys. Rev. Lett.* **58**, 203–206 (1987).
- Basché, Th., Moerner, W. E., Orrit, M. & Talon, H. Photon antibunching in the fluorescence of a single dye molecule trapped in a solid. *Phys. Rev. Lett.* **69**, 1516–1519 (1992).
- Fleury, L., Segura, J.-M., Zumofen, G., Hecht, B., Wild, U. P. Nonclassical photon statistics in single-molecule fluorescence at room temperature. *Phys. Rev. Lett.* **84**, 1148–1151 (2000).
- Nirmal, M. *et al.* Fluorescence intermittency in single cadmium selenide nanocrystals. *Nature* **383**, 802–804 (1996).
- Empedocles, S. & Bawendi, M. G. Spectroscopy of single CdSe nanocrystallites. *Acc. Chem. Res.* **32**, 389–396 (1999).
- Klimov, V. I. & McBranch, D. W. Femtosecond 1P-to-1S electron relaxation in strongly confined semiconductor nanocrystals. *Phys. Rev. Lett.* **98**, 4028–4031 (1998).
- Efros, A. L. & Rosen, M. Random telegraph signal in the photoluminescence intensity of a single quantum dot. *Phys. Rev. Lett.* **78**, 1110–1113 (1997).
- Murray, C. B., Norris, D. J. & Bawendi, M. G. Synthesis and characterization of nearly monodisperse CdE (E = S, Se, Te) semiconductor nanocrystallites. *J. Am. Chem. Soc.* **115**, 8706–8715 (1993).
- Hines, M. A. & Guyot-Sionnest, P. Synthesis and characterization of strongly luminescing ZnS-capped CdSe nanocrystals. *J. Phys. Chem.* **100**, 468–571 (1996).
- Dabbousi, B. O. *et al.* (CdSe)ZnS core shell quantum dots: synthesis and characterization of a size series of highly luminescent nanocrystallites. *J. Phys. Chem.* **101**, 9463–9475 (1997).
- Nirmal, M. & Brus, L. Luminescence photophysics in semiconductor nanocrystals. *Acc. Chem. Res.* **32**, 407–414 (1999).

Acknowledgements

This work is supported by the David and Lucile Packard Fellowship (A.I. and S.K.B.). P.M. acknowledges the financial support of the Max Kade Foundation.

Correspondence and requests for materials should be addressed to A.I. (e-mail: atac@ece.ucsb.edu).

Organoplatinum crystals for gas-triggered switches

Martin Albrecht[†], Martin Lutz[†], Anthony L. Spek[†] & Gerard van Koten^{*‡}

^{*}Debye Institute, Department of Metal-Mediated Synthesis & [†]Bijvoet Center for Biomolecular Research, Crystal and Structural Chemistry, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands

Considerable effort is being devoted to the fabrication of nano-scale devices¹. Molecular machines, motors and switches have been made, generally operating in solution^{2–7}, but for most device applications (such as electronics and opto-electronics), a maximal degree of order and regularity is required⁸. Crystalline materials would be excellent systems for these purposes, as crystals comprise a vast number of self-assembled molecules, with a perfectly ordered three-dimensional structure⁹. In non-porous crystals, however, the molecules are densely packed and any change in them (due, for example, to a reaction) is likely to destroy the crystal and its properties. Here we report the controlled and fully reversible crystalline-state reaction of gaseous SO₂ with non-porous crystalline materials consisting of organoplatinum molecules. This process, including repetitive expansion–reduction sequences (on gas uptake and release) of the crystal lattice, modifies the structures of these molecules without affecting their crystallinity. The process is based on the incorporation of SO₂ into the colourless crystals and its subsequent liberation from

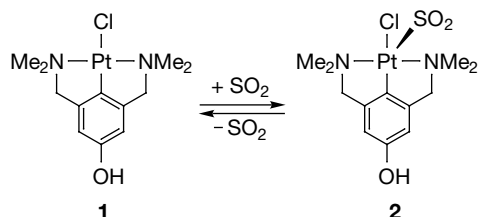


Figure 1 Reversible adsorption of SO_2 by organoplatinum(II) species containing the N,C,N' terdentate coordinating, monoanionic 'pincer' ligand.

the orange adducts by reversible bond formation and cleavage¹⁰. We therefore expect that these crystalline materials will find applications for gas storage devices and as opto-electronic switches^{11,12}.

Chemical transformations that occur in crystalline material, so-called 'crystalline-state reactions'¹³, are very rare because most chemical reactions cause considerable stress and intermolecular reorganization. Hence, the loss of crystallinity is a typical event. Among the best explored crystalline-state reactions are (reversible) photochemical isomerization processes, that is, unimolecular, light-induced transformations in the crystal^{14–16}. Processes in the crystalline phase that involve substrate binding and release, and thus a change in the overall atom content within the unit cell, have been

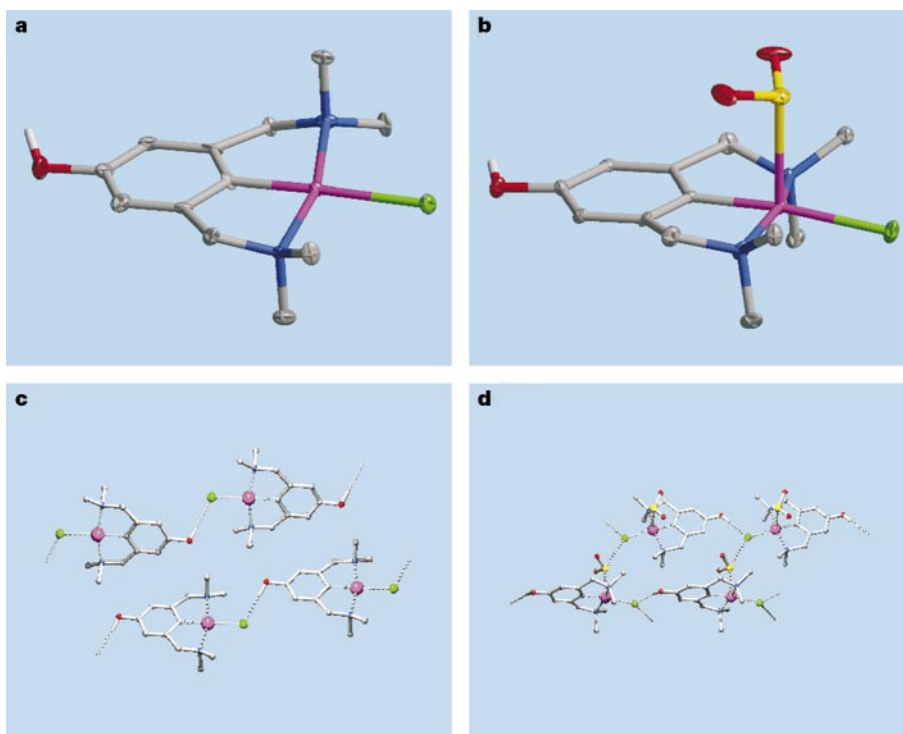


Figure 2 Molecular structures and hydrogen bonding motifs of organoplatinum complexes **1** and **2**. **a, b**, The molecular structures were determined by single crystal X-ray diffraction and demonstrate the change in the geometry around platinum upon coordination of SO_2 : distorted square-planar in **1** (**a**) versus square-pyramidal in the adduct **2** (**b**). **c, d**, The α -type supramolecular connectivity pattern, that is, the $\text{Pt}-\text{Cl}\cdots\text{H}-\text{O}$ hydrogen bonding, is similarly present in **1** (**c**) and in **2** (**d**). The β -type network mediated by $\text{Pt}\cdots\text{S}\cdots\text{Cl}$ interactions is only found in **2**. All hydrogen atoms except the phenolic $\text{O}-\text{H}$ hydrogen have been omitted for clarity. Elements are colour-coded: platinum (purple), nitrogen (blue), chlorine (green), oxygen (red), sulphur (yellow). X-ray crystallographic details for the single crystal structure determination for **1**: intensities were measured at 150 K using graphite-monochromated Mo $K\alpha$ radiation ($\lambda = 0.71073 \text{ \AA}$). 2,936 independent reflections (6,631 total measured) were analysed by using automated Patterson methods. Structure refinement on F^2 structure factors converged with a discrepancy factor $R_1 = 0.0376$ and a weighted discrepancy factor $wR_2 = 0.0793$

(goodness of fit = 1.024). The oxygen-bound hydrogen was located in the difference Fourier map and allowed to refine freely. Residual electron density was only found within 0.84 and $-1.08 \text{ e \AA}^{-3}$. An identical experimental set-up was used for the analysis of the adduct **2**. The structure was solved by using automated Patterson methods on 2,166 independent reflections (11,626 total measured). Structure refinement on F^2 converged at $R_1 = 0.0350$ and $wR_2 = 0.0558$ (goodness of fit = 1.054). The oxygen-bound hydrogen was located in the difference Fourier map and refined with a rotating model. Residual electron density was only found within 0.97 and $-1.01 \text{ e \AA}^{-3}$ (for further details, see refs 10 and 24, respectively). Selected bond distances (in $\text{\AA}$) and angles (in degrees) around platinum for **1**: $\text{Pt}-\text{C}$, 1.915(9); $\text{Pt}-\text{Cl}$, 2.434(2); $\text{Pt}-\text{N1}$, 2.094(8); $\text{Pt}-\text{N2}$, 2.082(8); $\text{N1}-\text{Pt}-\text{N2}$, 163.9(3); $\text{C}-\text{Pt}-\text{Cl}$, 177.4(4); and for **2**: $\text{Pt}-\text{C}$, 1.923(10); $\text{Pt}-\text{Cl}$, 2.423(3); $\text{Pt}-\text{N1}$, 2.106(8); $\text{Pt}-\text{N2}$, 2.096(8); $\text{N1}-\text{Pt}-\text{N2}$, 160.5(3); $\text{C}-\text{Pt}-\text{Cl}$, 173.1(3). (N1 and N2 are the respective nitrogen atoms, coloured blue).

Table 1 Selected crystallographic parameters of **1** and **2**

Complex	Space group (number)	Unit cell dimensions				Density, ρ (g cm^{-3})	Packing index (%)
		a ($\text{\AA}$)	b ($\text{\AA}$)	c ($\text{\AA}$)	V ($\text{\AA}^3$)		
1 (ref. 24)	$Pna2_1$ (33)	24.2238(14)	10.1986(8)	5.4483(14)	1,346.0(4)	2.1606(6)	72.2
2 (ref. 10)	$Pna2_1$ (33)	16.837(3)	10.2189(16)	9.0231(10)	1,552.5(4)	2.1473(6)	70.9

restricted to enzymes and other materials with an 'open' (that is, porous)¹⁷ structure with large cavities (for example, zeolites)^{18–20}. A versatile way of overcoming this problem relies on crystal engineering²¹, whose intention is the design and processing of supramolecular properties in solids²².

Our approach towards the property-directed synthesis of new and well-defined materials involves the use of the *N,C,N'* terdentate coordinating, monoanionic 'pincer' ligand system [C₆H₂(CH₂NMe₂)₂-2,6-R-4], abbreviated as NCN-R, where R is a functional group²³. When complexed to four-coordinated metal centres (for example, Pt(II)), the fourth site is forced to occupy a *trans* position to the metal–carbon bond, due to the chelating binding mode of the pincer ligand (Fig. 1). In addition, modification of the aryl ring of the ligand allows for the introduction of functional groups such as an acceptor or donor site. Using this method, the organoplatinum complex [PtCl(NCN-OH)] (**1**) was prepared recently²⁴; it contains a metal-bound chloride as a hydrogen bond acceptor and a phenolic hydroxide group on the pincer ligand as a hydrogen donor. With this supramolecular synthon²⁵, the organoplatinum complex self-assembles in the solid state to form an α -type network via intermolecular Pt–Cl...H–O (hydrogen) bonds (Fig. 2).

When crystalline **1** is exposed to an atmosphere of SO₂, adsorption of this gas by the organoplatinum sites is indicated by a dramatic colour change of the material from colourless to deep orange, a colour change previously also observed in solution^{10,26}. This adduct formation modifies the geometry around platinum and was therefore expected to destroy the crystallinity. Surprisingly however, exposure of crystalline **1** to SO₂ gas (about 1 minute exposure time) gives the corresponding adduct [PtCl(NCN-OH)(SO₂)]₂, **2** (Fig. 1), in a quantitative, crystalline-state reaction. Unequivocal evidence for this process has been obtained by X-ray powder diffraction and

solid-state infrared spectroscopy. Similarly, the reverse reaction in an SO₂-free environment leads to the complete regeneration of crystalline **1**.

The structural features of **1** and **2** are known from single crystal analyses taken on crystals that were grown from saturated solutions (Fig. 2a, b)^{10,24}. On absorption of SO₂, the molecular structure is changed predominantly around the platinum atom. The metal centre is no longer of square-planar but of approximate square-pyramidal geometry. This has an influence on the packing index and the density, which decrease, and on the unit cell volume, which is expanded by more than 15% after absorption of SO₂ (Table 1). Intermolecular Pt–Cl...H–O hydrogen bonding in **1** results in the formation of an α -type network (Fig. 2c). Tight fixation of SO₂ in **2** occurs by a Pt...S(O)₂...Cl bond formation perpendicular to the hydrogen bond network²⁷, which results in a unusual β -type network (Fig. 2d). The distances between hydrogen bond donors and acceptors (H...Cl is 2.33(9) Å in **2** and 2.32(13) Å in **1**; O...Cl is 3.127(8) Å in **2** and 3.126(8) Å in **1**) as well as the bond angles (O–H...Cl is 165(8)° in **2** and 161(15)° in **1**) are statistically identical.

When SO₂-free organoplatinum crystals are subjected to an environment of SO₂ for a few seconds, only partial binding of the substrate is observed and a mixture of the two crystalline compounds **1** and **2** is obtained. The precise composition of this mixture is strongly dependent on the exposure time and the sample preparation. However, the powder diffraction diagram of the mixture can be used to access the ratio of **1** and **2** because their patterns are very different. Therefore, the comparison of the measured powder diagram with those simulated from the single crystal structures affords a good estimation of the ratio between **1** and **2**.

The reversibility and kinetics of these crystalline-state reactions have been elucidated by time-resolved X-ray powder diffraction²⁸ and infrared spectroscopy. By overlapping of the time-resolved infrared spectra, various isosbestic points have been identified which strongly suggest a direct transformation of **1** into **2** and vice versa. We examined the stretching vibrations of bound SO₂ ($\nu_s = 1,072$ and $\nu_{as} = 1,236$ cm^{–1}). The (dis)appearance of these

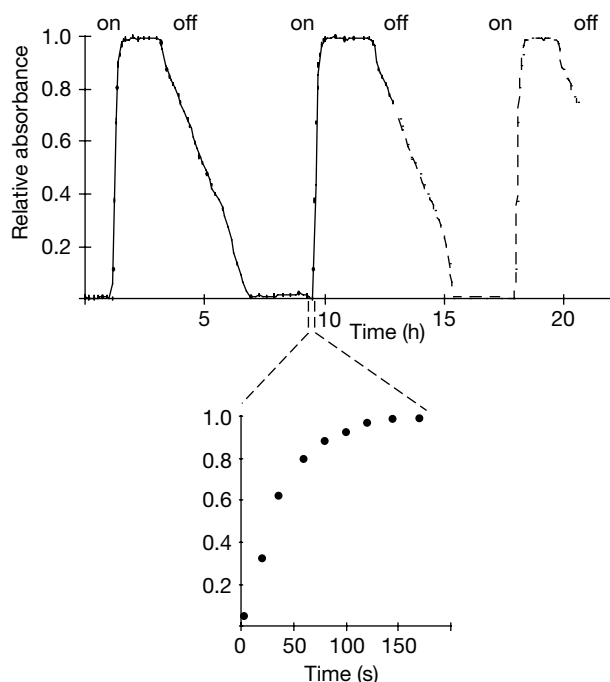


Figure 3 'On/off' switching between crystalline **1** and **2**. Switching was monitored by infrared spectroscopy in an environmental DRIFT (diffuse reflectance infrared Fourier transform) chamber. The vibration of platinum-bound SO₂ ($\nu_s = 1,072$ cm^{–1}) is diagnostic for the position of the switch. Spectra were recorded at 293 K every 20 seconds (average of 4 scans) under a continuous flow of SO₂ (crystalline-state reaction giving **2**) and N₂ (giving **1**), respectively. Repetitive switching 'on' and 'off' does not reduce the amplitude.

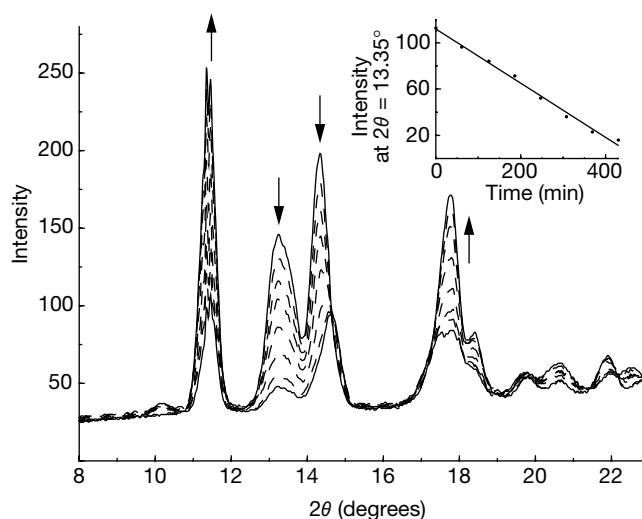


Figure 4 Time-resolved X-ray powder diffraction analysis of the transformation of crystalline **2** to **1** in an atmosphere of air. **2** was mounted in an open glass capillary on an Nonius Kappa charge-coupled device (CCD) diffractometer with Cu K α radiation, graphite monochromator, $\lambda = 1.5418$ Å at a temperature of 293 K. Scans were performed every hour with a scan time of 6 minutes. After background subtraction, the characteristic intensities at $2\theta = 11.33$ (diagnostic for **1**) and 13.35 (for **2**) have been related to the reaction time. Inset, correlation exceeds 0.996 for a linear approximation.

signals is dependent on the atmospheric constitution and thus demonstrates the potential of this material as a crystalline switch (Fig. 3). The gas fixation process of crystalline **1** ('off' position) in a steady atmosphere of SO₂ is best approximated by a pseudo first-order rate law in organoplatinum species (that is, rate = $k[\text{Pt}]$ where $[\text{Pt}]$ is the concentration of **1**; Fig. 3), whereas a linear decrease of the characteristic absorption bands is noted when **2** ('on' position) is exposed to an inert gas. (In fact, k is (to our knowledge) the first reliable rate constant measured for the diffusion of a gas through a crystalline material. It is strongly temperature dependent: at a given particle size, the gas release is accelerated approximately fivefold when the temperature is raised from ambient (293 K) to 333 K and enormously retarded at low temperature (253 K, about 50 times slower). Preliminary measurements suggest that the gas desorption mechanism at higher temperatures may be different from that operating at ambient temperatures.) This suggests that in the rate-determining step of substrate binding, single platinum sites are involved. Although absorption of SO₂ on molecules which are located at the surface of the crystalline material is assumed to be facile, the transport of the gas into the interior of the crystal to neighbouring molecules is probably less favoured. This can be explained by considering competitive release of SO₂ back to the atmosphere. More importantly, incorporation of the substrate into the inner sphere is considerably complicated as an expansion of the crystal lattice occurs, involving, for example, reorganization of the relative position of the α -type networks (Fig. 2). This implies that building blocks containing bound SO₂ must influence their surrounding substrate-free neighbours in such a way that space is available for the gas molecule to move inside the densely packed supermolecule.

Using powder diffraction techniques, gradual transformation of **2** to **1** without formation of an intermediate has been observed, and the decrease of the intensity peaks of crystalline adduct **2** is accompanied by a proportional increase of the intensities assigned to **1** (see above, Fig. 4). Analysis of the characteristic intensities at low 2θ values strongly suggests a linear relationship between the change of the intensities and the elapsed time (see inset to Fig. 4), thus corroborating the results obtained from infrared spectroscopy (see above). Hence, a zero-order rate law is deduced for the release of SO₂ gas from crystalline **2**. The exact reaction rate is strongly dependent on the surface area of the sample (for example, particle size). Moreover, in a thin capillary, the desorption reaction of SO₂ from **2** proceeds considerably more slowly than in an open cup, but is still linear with time. This result emphasises the importance of area and the properties of the surface with respect to exchange with

the local environment. Additionally, single crystals of **2** lose SO₂ primarily at the solid–gas interphase, resulting in the observation that a colourless zone 'grows' from the outside to the inside of the crystal block (Fig. 5). The desorption of SO₂ at the surface is assumed to be relatively fast and therefore the rate-determining step is presumably the diffusion-controlled transportation of the gaseous molecule from one metal centre to another, that is, the transfer of SO₂ from the core to the surface of the crystal.

The full reversibility of the crystalline-state absorption and desorption reaction is demonstrated by exposing a sample of regenerated **1** to atmospheres of air and SO₂ alternately. No loss in signal intensities, for example, owing to the formation of amorphous powder, is observed even after several repeated cycles. When single crystals of **2** are exposed to air, **1** is obtained as a colourless crystalline powder that is no longer suitable for a full single crystal X-ray analysis.

These results provide access to the preparation and engineering of efficient crystalline switches with unusual properties. The 'on/off' position of this switch may be defined, for example, by the crystal size and properties, or by the diagnostic colour of these materials. Most importantly, such switches are triggered primarily by their gaseous environment and may also be regulated by temperature. Owing to the insensitivity of the present organoplatinum systems towards light or oxidation^{13–16}, these materials have high potential for applications as gas sensors²⁹ and as photochemical switches in opto-electronic devices in, for example, data processing or non-linear optical technology. □

Received 19 April; accepted 21 June 2000.

1. Fox, M. A. Fundamentals in the design of molecular electronic devices: long-range charge carrier transport and electronic coupling. *Acc. Chem. Res.* **32**, 201–207 (1999).
2. Kelly, T. R., De Silva, H. & Silva, R. A. Unidirectional rotary motion in a molecular system. *Nature* **401**, 150–152 (1999).
3. Koumura, N., Zijlstra, R. W. J., van Delden, R. A., Harada, N. & Feringa, B. L. Light-driven monodirectional molecular rotor. *Nature* **401**, 152–155 (1999).
4. Balzani, V., Gómez-López, M. & Stoddart, J. F. Molecular machines. *Acc. Chem. Res.* **31**, 405–414 (1998).
5. Sauvage, J. P. Transition metal-containing rotaxanes and catenanes in motion: toward molecular machines and motors. *Acc. Chem. Res.* **31**, 611–619 (1998).
6. Fabbrizzi, L., Licchelli, M. & Pallavicini, P. Transition metals as switches. *Acc. Chem. Res.* **32**, 846–853 (1999).
7. Collier, C. P. *et al.* Electronically configurable molecular-based logic gates. *Science* **285**, 391–394 (1999).
8. Liu, C.-Y. & Bard, A. J. Optoelectronic properties and memories based on organic single-crystal thin films. *Acc. Chem. Res.* **32**, 235–245 (1999).
9. Dunitz, J. D. in *Perspectives in Supramolecular Chemistry. The Crystal as a Supramolecular Entity* (ed. Desiraju, G. R.) **7** (Wiley, Chichester, 1996).
10. Albrecht, M., Gossage, R. A., Lutz, M., Spek, A. L. & van Koten, G. Diagnostic organometallic and metalloendritic materials for SO₂ gas detection: reversible binding of sulfur dioxide to aryl platinum(II) complexes. *Chem. Eur. J.* **6**, 1431–1445 (2000).
11. Chen, P., Wu, X., Liu, J. & Tan, K. L. High H₂ uptake by alkali-doped carbon nanotubes under ambient pressure and moderate temperatures. *Science* **285**, 91–93 (1999).
12. Dillon, A. C. *et al.* Storage of hydrogen in single-walled carbon nanotubes. *Nature* **386**, 377–379 (1997).
13. Ohashi, Y., Yanagi, K., Kurihara, T., Sasada, Y. & Ohgo, Y. Crystalline-state reaction of cobaloxime complexes by X-ray exposure. An order-to-order racemization in the crystal of [(S)-1-cyanoethyl](pyridine)-bis(dimethylglyoximate)cobalt(III). *J. Am. Chem. Soc.* **104**, 6353–6359 (1982).
14. Novak, K., Enkelmann, V., Wegner, G. & Wagener, K. B. Crystallographic study of a single crystal to single crystal photodimerization and its thermal reverse reaction. *Angew. Chem. Int. Edn Engl.* **32**, 1614–1616 (1993).
15. Kobatake, S., Yamada, T., Uchida, K., Kato, N. & Irie, M. Photochromism of 1,2-bis(2,5-dimethyl-3-thienyl)perfluorocyclopentene in a single crystalline phase. *J. Am. Chem. Soc.* **121**, 2380–2386 (1999).
16. Scheffer, J. R. & Pokkuluri, P. R. in *Photochemistry in Organized & Constrained Media* (ed. Ramamurthy, V.) **185** (VCH, New York, 1990).
17. Langeley, P. J. & Hulliger, J. Nanoporous and mesoporous organic structures: new openings for materials research. *Chem. Soc. Rev.* **28**, 279–291 (1999).
18. Estermann, M., McCusker, L. B., Baerlocher, C., Merouche, A. & Kessler, H. A synthetic gallo phosphate molecular sieve with a 20-tetrahedral-atom pore opening. *Nature* **352**, 320–323 (1991).
19. Hajdu, J. *et al.* Millisecond X-ray diffraction and the first electron density map from Laue photographs of a protein crystal. *Nature* **329**, 178–181 (1987).
20. Buss, C. E. *et al.* Structural investigations of vapochemical behavior. X-ray single-crystal and powder diffraction studies of [Pt(CN-iso-C₃H₇)₄][M(CN)₄] for M = Pt or Pd. *J. Am. Chem. Soc.* **120**, 7783–7790 (1998).
21. Desiraju, G. R. *Crystal Engineering: The Design of Organic Solids* (Elsevier, Amsterdam, 1989).
22. Lehn, J.-M. *Supramolecular Chemistry: Concepts and Perspectives* (VCH, Weinheim, 1995).
23. van Koten, G. Tuning the reactivity of metals held in a rigid ligand environment. *Pure Appl. Chem.* **61**, 1681–1694 (1989).

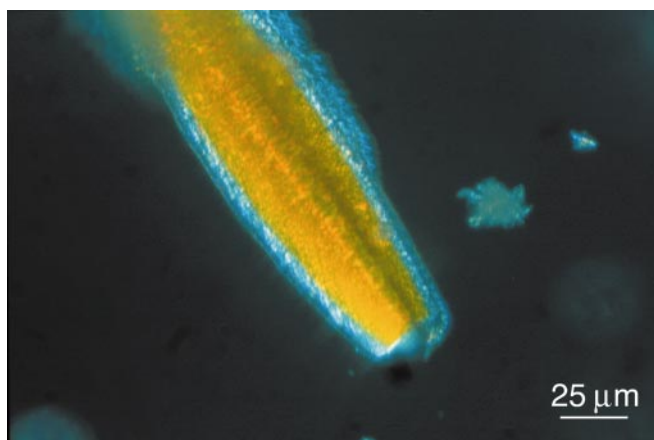


Figure 5 Formation of crystalline **1** from a single crystal of **2**. Shown after about 14 h in air, showing the colourless zone at the periphery and the orange colour in the core of the crystal.

24. Davies, P. J. *et al.* Organoplatinum building blocks for one-dimensional hydrogen-bonded polymeric structures. *Angew. Chem. Int. Edn Engl.* **35**, 1959–1961 (1996).
25. Braga, D., Grepioni, F. & Desiraju, G. R. Crystal engineering and organometallic architecture. *Chem. Rev.* **98**, 1375–1405 (1998).
26. Albrecht, M. & van Koten, G. Gas sensor materials based on metalodendrimers. *Adv. Mater.* **11**, 171–174 (1999).
27. Darenbourg, M. Y., Tuntulani, T. & Reibenspies, J. H. Structure/function relationships in ligand-based SO₂/O₂ conversion to sulfate as promoted by nickel and palladium thiolates. *Inorg. Chem.* **34**, 6287–6290 (1995).
28. Moffat, K. Time-resolved crystallography. *Acta Crystallogr. A* **54**, 833–841 (1998).
29. Kong, J. *et al.* Nanotube molecular wires as chemical sensors. *Science* **287**, 622–625 (2000).

Acknowledgements

We thank E. T. H. Lutz and A. M. M. Schreurs for technical assistance during the measurements and R. A. Gossage for discussions. This work was partially supported by the Council for Chemical Sciences from the Dutch Organization for Scientific Research (CW-NWO).

Correspondence and requests for material should be addressed to G.v.K. (e-mail: g.vankoten@chem.uu.nl).

Automatic design and manufacture of robotic lifeforms

Hod Lipson & Jordan B. Pollack

Computer Science Department, Volen Center for Complex Systems, Brandeis University, Waltham, Massachusetts 02454, USA

Biological life is in control of its own means of reproduction, which generally involves complex, autocatalysing chemical reactions. But this autonomy of design and manufacture has not yet been realized artificially¹. Robots are still laboriously designed and constructed by teams of human engineers, usually at considerable expense. Few robots are available because these costs must be absorbed through mass production, which is justified only for toys, weapons and industrial systems such as automatic teller machines. Here we report the results of a combined computational and experimental approach in which simple electromechanical systems are evolved through simulations from basic building blocks (bars, actuators and artificial neurons); the 'fittest' machines (defined by their locomotive ability) are then fabricated robotically using rapid manufacturing technology. We thus achieve autonomy of design and construction using evolution in a 'limited universe' physical simulation^{2,3} coupled to automatic fabrication.

In the field of artificial life, 'life as it could be' is examined on the basis of understanding the principles, and simulating the mechanisms, of real biological forms⁴. Just as aeroplanes use the same principles as birds, but have fixed wings, artificial lifeforms may share the same principles, but not the same implementation in chemistry. Stored energy, autonomous movement, and even animal communication are replicated in toys using batteries, motors and computer chips.

Our central claim is that to realize artificial life, full autonomy must be attained not only at the level of power and behaviour (the goal of robotics, today⁵), but also at the levels of design and fabrication. Only then can we expect synthetic creatures to sustain their own evolution. We thus seek automatically designed and constructed physical artefacts that are functional in the real world, diverse in architecture (possibly each slightly different), and automatically producible with short turnaround time, at low cost and in large quantities. So far these requirements have not been met.

The experiments described here use evolutionary computation for design, and additive fabrication for reproduction. The evolutionary process operates on a population of candidate robots,

each composed of some repertoire of building blocks. The evolutionary process iteratively selects fitter machines, creates offspring by adding, modifying and removing building blocks using a set of operators, and replaces them into the population (see Methods section). Evolutionary computation has been applied to many engineering problems⁶. However, studies in the field of evolutionary robotics reported to date involve either entirely virtual worlds^{2,3}, or, when applied in reality, adaptation of only the control level of manually designed and constructed robots^{7–9}. These robots have a predominantly fixed architecture, although Lund¹⁰ evolved partial aspects of the morphology, Thompson¹¹ evolved physical electric circuits for control only, and we evolved static Lego structures, but had to manually construct the resultant designs¹². Other works involving real robots make use of high-level building blocks comprising significant pre-programmed knowledge¹³. Similarly, additive fabrication technology has been developing in terms of materials and mechanical fidelity¹⁴ but has not been placed under the control of an evolutionary process.

Our approach is based on the use of only elementary building blocks and operators in both the design and fabrication process. As building blocks are more elementary, any inductive bias associated with them is minimized, and at the same time architectural flexibility is maximized. Similarly, use of elementary building blocks in the fabrication process allows the latter to be more systematic and versatile. As a theoretical extreme, if we could use only atoms as building blocks, laws of physics as constraints and nanomanipulation for fabrication, the versatility of the manufacturable design space would be maximized. Earlier reported work that used higher-level components and limited architectures (such as only tree structures^{2,3}) resulted in expedited convergence to acceptable solutions, but at the expense of truncating the design space. Furthermore, these design spaces did not consider manufacturability.

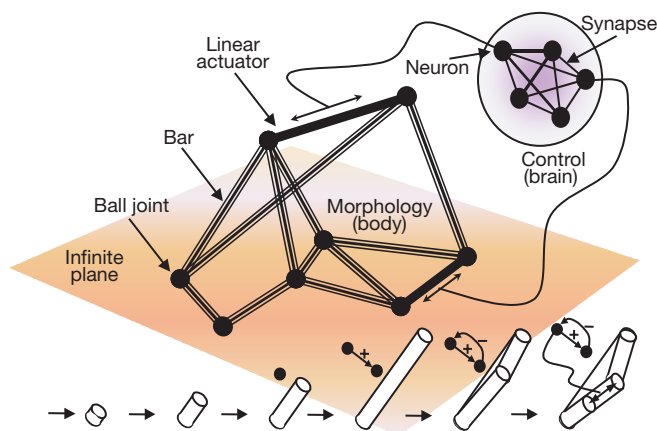


Figure 1 Schematic illustration of an evolvable robot. Bars connect to each other to form arbitrary trusses; by changing the number of bars and the way they connect, the structural behaviour of the truss is modified—some substructures may become rigid, while others may become articulated. Neurons connect to each other via synapses to form arbitrary recurrent neural networks. By changing the synapse weights and the activation threshold of the neuron, the behaviour of the neuron is modified. By changing the number of neurons and their connectivity, the behaviour of the network is modified. Also, we allow neurons to connect to bars: in the same way that a real neuron governs the contraction of muscle tissue, the artificial neuron signal will control the length of the bar by means of a linear actuator. All these changes can be brought about by mutational operators. A sequence of operators will construct a robot and its controller from scratch by adding, modifying and removing building blocks. The sequence at the bottom of the image illustrates an arbitrary progression of operators that create a small bar, elongate it and split it. Simultaneously, other operators create a neuron, add another neuron, connect them in a loop, and eventually connect one of the neurons to one of the bars. The bar is now an actuator. Because no sensors were used, these robots can only generate patterns and actions, but cannot directly react to their environment.

The design space that we used consisted of bars and actuators as building blocks of structure and artificial neurons as building blocks of control. Bars connected with free joints can potentially form trusses—that can represent arbitrary rigid, flexible and articulated structures—as well as multiple detached structures, and emulate revolving, linear and planar joints at various levels of hierarchy. Similarly, artificial neurons can connect to create arbitrary control architectures such as feed-forward and recurrent nets, state machines and multiple independent controllers (like multiple ganglia). Additive fabrication, where structure is generated layer by layer, allows for the automatic generation of arbitrarily complex physical structures and the rapid construction of physically different bodies, including any that are composed of our building blocks. A schematic illustration of a possible architecture is shown in Fig. 1. The bars connect to each other through ball-and-socket joints, neurons can connect to other neurons through synaptic connections, and neurons can connect to bars. In the latter case, the length of the bar is governed by the output of the neuron by means of a linear actuator. No sensors were used.

Starting with a population of 200 machines that were composed initially of zero bars and zero neurons, we conducted evolution in simulation. The fitness of a machine was determined by its locomotion ability—the net distance that its centre of mass moved on an infinite plane in a fixed duration. The process iteratively selected fitter machines, created offspring by adding, modifying and removing building blocks, and replaced them into the population (see

Methods). This process typically continued for 300 to 600 generations. Both body (morphology) and brain (control) were thus co-evolved simultaneously.

The simulator that we used for evaluating fitness (see Methods) supported quasi-static motion in which each frame is statically stable. This kind of motion is simpler to transfer reliably into reality, yet is rich enough to support low-momentum locomotion. Typically, several tens of generations passed before the first movement occurred. For example, at a minimum, a neural network generating varying output must assemble and connect to an actuator for any motion at all (see the sequence in Fig. 1 for an example). Various patterns of evolutionary dynamics emerged, some of which are reminiscent of natural phylogenetic trees. Figure 2 presents examples of extreme cases of convergence, speciation and massive extinction. A sample instance of an entire generation thinned down to unique individuals is shown in Fig. 3.

Selected (virtual) robots out of those with winning performance were then automatically converted into physical objects: their bodies, represented only as points and lines, were first expanded into solid models with ball joints and accommodations for linear motors according to the evolved designs (Fig. 4a). This 'solidifying' stage was performed by an automatic program that combined pre-designed components describing a generic bar, ball joint, and actuator. The virtual solid bodies were then 'materialized' using commercial rapid prototyping technology (Fig. 4b). This machine used a temperature-controlled head to extrude thermoplastic

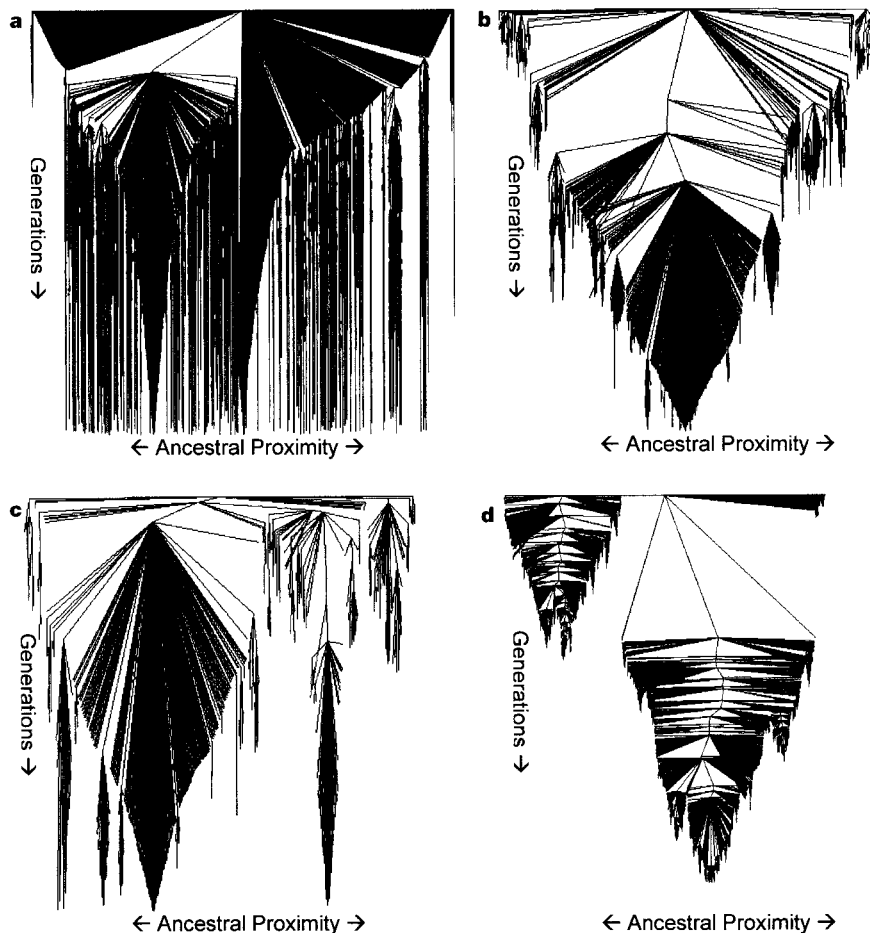


Figure 2 Phylogenetic trees of several different evolutionary runs. Each node in the tree represents an individual and links represent parent–child relationships. The vertical axis represents generations, and the horizontal axis represents ancestral proximity in terms of the hops along the tree necessary to get from one individual to another. All trees originate at a common root denoting an empty robot with zero bars and actuators. Trees exhibit

various degrees of divergence and speciation: **a**, extreme divergence, resulting from niching methods²²; **b**, extreme convergence, resulting from fitness-proportionate selection; **c**, intermediate level of divergence, typical of earlier stages of fitness-proportionate selection; and **d**, massive extinction under fitness-proportionate selection. The trees are thinned, and depict several hundred generations each.

material layer by layer, so that the arbitrarily evolved morphology emerged as a solid three-dimensional structure without tooling or human intervention. The entire pre-assembled machine was fabricated as a single unit, with fine plastic supports connecting between moving parts (Fig. 4c); these supports broke away at first motion. The resulting structures contained complex joints that would be difficult to design or manufacture using traditional methods (Figs 4d and 5). Standard stepper motors were then snapped in, and the evolved neural network was executed on a microcontroller to activate the motors. The physical machines (three to date) then faithfully reproduced their virtual ancestors' behaviour in reality (see Table 1).

In spite of the relatively simple task and environment (locomotion over an infinite horizontal plane), surprisingly different and elaborate solutions were evolved. Machines typically contained around 20 building blocks, sometimes with significant redundancy (perhaps to make mutation less likely to be catastrophic¹⁵). Not less surprising was the fact that some (for example, Fig. 5b) exhibited symmetry, which was neither specified nor rewarded for anywhere in the code; a possible explanation is that symmetric machines are more likely to move in a straight line, consequently covering a greater net distance and acquiring more fitness. Similarly, successful designs appear to be robust in the sense that changes to bar lengths

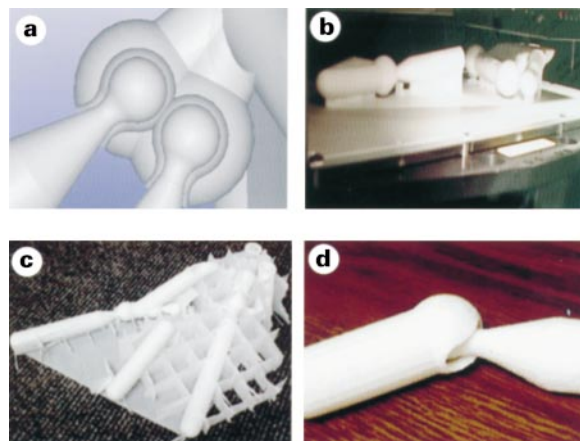


Figure 4 Physical embodiment process. **a**, Automatically 'fleshed' joints in virtual space; **b**, a physical replication process in a rapid prototyping machine that builds the three-dimensional morphology layer after layer; **c**, pre-assembled body in mid print with discardable support structure; **d**, a close-up image of a joint printed as a single unit. The ball is printed inside the socket.

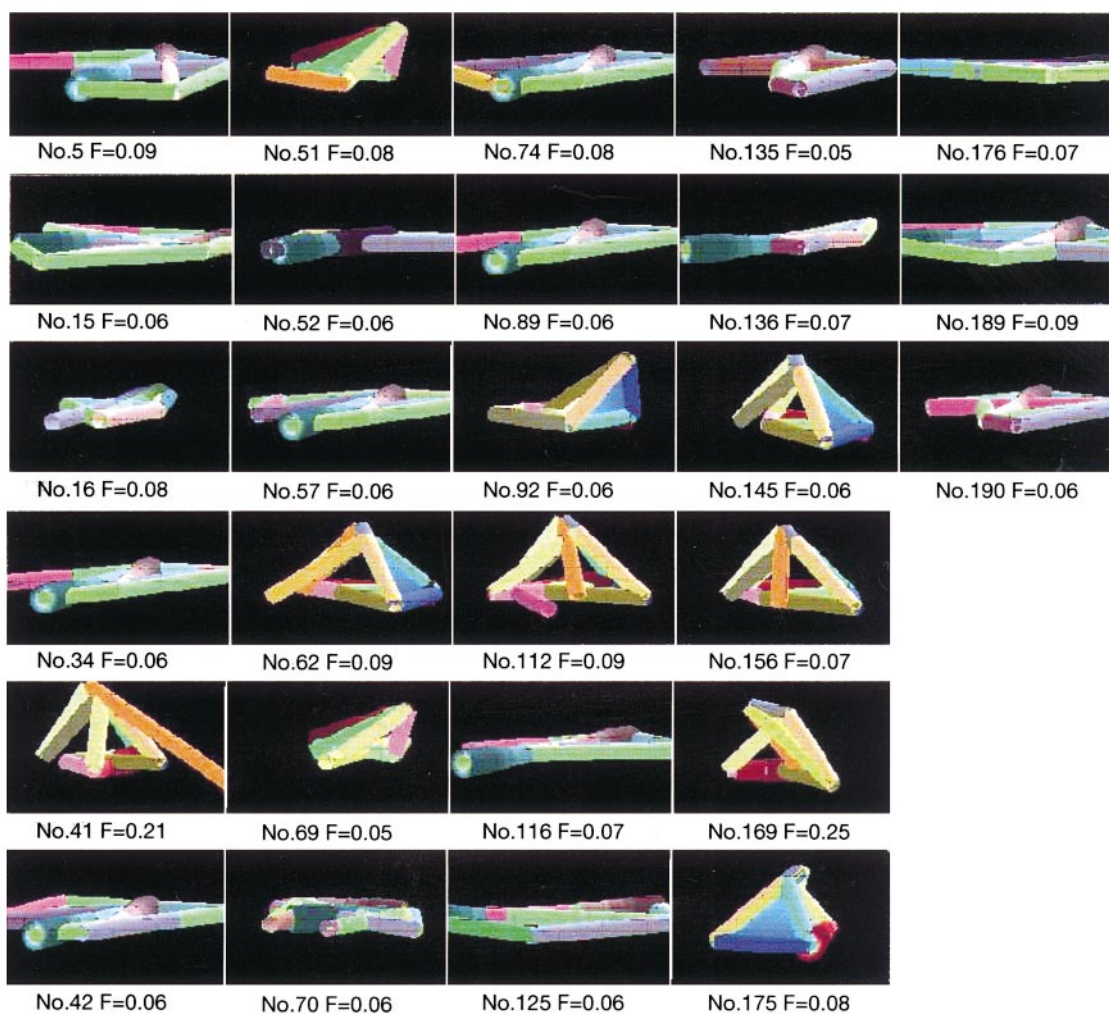


Figure 3 A generation of robots. An arbitrarily sampled instance of an entire generation, thinned down to show only significantly different individuals. The caption under each image provides an arbitrary index number (used for reference) and the fitness of that

individual. Two subpopulations of robots are observable, each with its own variations: one flat on the ground, and the other containing some elevated structure.

would not significantly hamper their mobility. Three samples are shown and described in detail in Fig. 5, exploiting principles of ratcheting (Fig. 5a), anti-phase synchronization (Fig. 5b) and dragging (Fig. 5c). Others (not shown) used a sort of a crawling bipedalism, where a body resting on the floor is advanced using alternating thrusts of left and right 'limbs'. Some mechanisms used sliding articulated components to produce crab-like sideways motion. Other machines used a balancing mechanism to shift a friction point from side to side and advance by oscillatory motion. Table 1 compares the performances of three physical machines to their virtual ancestors. We note that although overall distance travelled in the second and third cases does not match, in all cases the physical motion was achieved using corresponding mechanical and control principles. The difference in distance results from the limbs slipping on the surface, implying that the friction model used in the simulation was not realistic.

Although both the machines and the task that we describe here are fairly simple compared with the products of human teams of engineers (and with the products of biological evolution), we have

demonstrated a robotic 'bootstrap' process, in which automatically designed electromechanical systems have been manufactured robotically. We have carefully minimized human intervention in both the design and the fabrication stages. Apart from snapping in the motors, the only human work was in informing the simulation about the 'universe' that could be manufactured.

Without reference to specific organic chemistry, life is an autonomous design process that is in control of a complex set of chemical 'factories', allowing the generation and testing of physical entities that exploit the properties of the medium of their own construction. Using a different medium, namely off-the-shelf rapid manufacturing, and evolutionary design in simulation, we have made progress towards replicating this autonomy of design and manufacture. This is, to our knowledge, the first time any artificial evolution system has been connected to an automatic physical construction system. Taken together, our evolutionary design system, 'solidification' process, and rapid prototyping machine form a primitive replicating robot. Although there are many, many further steps before this technology could become dangerous¹⁶, we believe that if indeed artificial systems are to ultimately interact and integrate with reality, they cannot remain virtual; it is crucial that they cross the simulation–reality gap to learn, to evolve¹⁷, and to affect the physical world directly¹⁸. Eventually, the evolutionary process must accept feedback from the live performance of its products.

Future work is needed primarily in understanding how more complex modular structures might self-organize, and how these complex structures may transfer into reality under control of the evolutionary process. Technological advances in micro-electro-mechanic systems (MEMS), nanofabrication, and multi-material rapid prototyping that can embed circuits¹⁹ and actuators²⁰ in bulk material, together with higher-fidelity physical simulation and an increased understanding of evolutionary computational processes, may pave the way for the self-sustaining progress that Moravec has termed "escape velocity"²¹. □

Methods

Robot representation

A robot was represented by a string of integers and floating-point numbers that describe bars, neurons and their connectivity, as follows:

robot := (vertices)(bars)(neurons)(actuators)

vertex := (x, y, z)

bar := (vertex 1 index, vertex 2 index, relaxed length, stiffness)

neuron := (threshold, synapse coefficients of connections to all neurons)

actuator := (bar index, neuron index, bar range)

Evolution process

Experiments were performed using version 1.2 of GOLEM (Genetically Organized Lifelike Electro Mechanics), which is available at <http://www.demo.cs.brandeis.edu/golem>. We performed a simulated evolutionary process: the fitness function was defined as the net euclidean distance that the centre-of-mass of an individual moves over a fixed number (12) of cycles of its neural control. We started with a population of 200 null (empty) individuals. Each experiment used a different random seed. Individuals were then selected, mutated, and replaced into the population in steady state as follows: the selection functions we tried were random, fitness-proportionate or rank-proportionate. The mutation operators used to generate an offspring were independently applied with the following probabilities: a small mutation in length of bar or neuron synaptic weight (0.1), the removal or addition of a small dangling bar or unconnected neuron (0.01), split vertex into two and add a small bar, or split bar into two and add vertex (0.03), attach or detach neuron to bar (0.03). At least one mutation was applied. The mutations took place on the symbolic representation of the phenotype. After mutation, a new fitness was assigned to the individual by means of a simulation of the mechanics and the control (see details below). The offspring was inserted into the population by replacing an existing individual. The replacement functions we tried chose individuals to replace either randomly, in inverse-proportion to their fitness, or using similarity-proportionate criteria (deterministic crowding²²). Various permutations of selection-replacement methods are possible; the results we report here were obtained using fitness-proportionate selection and random replacement. However, using rank selection instead of fitness-proportionate selection, or using random selection with fitness-proportionate replacement yielded equivalent results. The process continued for 300 to 600 generations (approximately 10⁵ evaluations overall).

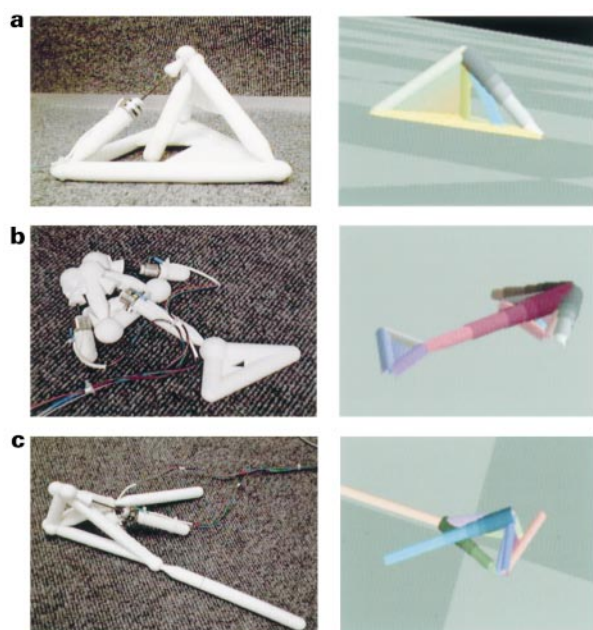


Figure 5 Three resulting robots. Real robots (left); simulated robots (right). **a**, A tetrahedral mechanism that produces hinge-like motion and advances by pushing the central bar against the floor. **b**, This surprisingly symmetric machine uses a seven-neuron network to drive the centre actuator in perfect anti-phase with the two synchronized side limb actuators. While the upper two limbs push, the central body is retracted, and vice versa. **c**, This mechanism has an elevated body, from which it pushes an actuator down directly onto the floor to create ratcheting motion. It has a few redundant bars dragged on the floor, which might be contributing to its stability. Print times are 22, 12 and 18 hours, respectively. These machines perform in reality in the same way that they perform in simulation. Motion videos of these robots and others are available: see Supplementary Information.

Table 1 Results

Machine	Distance travelled (cm)	
	Virtual	Physical
Tetrahedron (Fig. 5a)	38.5	38.4 (35)
Arrow (Fig. 5b)	59.6	22.5 (18)
Pusher (Fig. 5c)	85.1	23.4 (15)

Comparison of the performance of physical machines versus their virtual origin. Values are the net distance that the centre of mass of each machine travelled over 12 cycles of neural network. Distances given in the column headed 'physical' are compensated for scale reduction (actual distance is shown in parentheses). The mismatch in the last two rows is primarily due to the slipping of limbs on the surface.

The process was performed both serially and in parallel (on a 16-processor computer). On parallel computers we noticed an inherent bias towards simplicity: simpler machines could complete their evaluation sooner and consequently reproduce more quickly than complex machines (this could be avoided with a generational implementation).

Our evolutionary simulation was based on evolutionary strategies²³ and evolutionary programming²⁴, because it directly manipulated continuous valued representations and used only elementary operators of mutation. Alternatively, we could have used genetic algorithms²⁵ and genetic programming²⁶ that introduce crossover operators that are sensitive to the structure of the machines, which might change the rate of evolution and lead to replicated structures. We did not form a morphological grammar from which the body is developed²⁷, but evolved directly on the symbolic representation of the phenotype. And, instead of separating body (morphology) and brain (control) into separate populations, or providing for a 'neonatal' stage that might allow us to select for brains that are able to learn to control their bodies, we simply applied selection to bodies and brains as integrated units. This simplified experimental set-up followed our focus on completing the simulation and reality loop, but we anticipate that the many techniques that have been developed in evolutionary and co-evolutionary learning^{28–30} will enrich our results.

Simulation

Both the mechanics and the neural control of a machine were simulated concurrently. The mechanics were simulated using quasi-static motion, where each frame of the motion was assumed to be statically stable. This kind of motion is simple to simulate and easy to induce in reality, yet is rich enough to support various kinds of low-momentum motion like crawling and walking (but not jumping). The model consisted of ball-jointed cylindrical bars with true diameters. Each frame was solved by relaxation: an energy term was defined, taking into account elasticity of the bars, potential gravitational energy, and penetration energy of collision and contact. The degrees of freedom of the model (vertex coordinates) were then adjusted iteratively according to their derivatives to minimize the energy term, and the energy was recalculated. Static friction was also modelled. The use of relaxation permitted handling singularities (for example, snap-through buckling) and under-constrained cases (like a dangling bar). Noise was added to ensure the system does not converge to unstable equilibrium points, and to cover the simulation–reality gap. The material properties modelled correspond to the properties of the rapid prototyping material (modulus of elasticity, $E = 0.896$ GPa; specific density, $\rho = 1,000$ kg m⁻³; yield stress, $\sigma_{\text{yield}} = 19$ MPa). The neural network was simulated in discrete synchronized cycles. In each cycle, actuator lengths were modified in small increments not larger than 1 cm.

Received 6 March; accepted 27 June 2000.

- Langton, C. *Artificial Life* (Addison-Wesley, Redwood City, California, 1989).
- Sims, K. in *Proc. 4th Artificial Life Conf.* (eds Brooks, R. & Maes, P.) 28–35 (MIT Press, Cambridge, MA, 1994).
- Komosinski, M. & Ulatowski, S. in *5th Eur. Conf. on Artificial Life ECAL '99* (eds Floreano, D. et al.) 261–265 (Springer, Berlin, 1999).
- Smith, J. M. Byte-sized evolution. *Nature* **355**, 772–773 (1992).
- Swinson, M. *Mobile Autonomous Robot Software* (Report BAA-99-09, DARPA, Arlington, Virginia, 1998).
- Bentley, P. (ed.) *Evolutionary Design by Computers* (Morgan Kaufmann, San Francisco, 1999).
- Floreano, D. & Mondada, F. in *From Animals to Animats III* (eds Cliff, D., Husbands, P., Meyer, J. & Wilson, S.) 421–430 (MIT Press, Cambridge, MA, 1994).
- Husbands, P. & Meyer, J. A. *Evolutionary Robotics* (Springer, Berlin, 1998).
- Nolfi, S. Evolving non-trivial behaviors on real-robots: a garbage collecting robot. *Robotics and Autonomous Systems*, **22**, 187–198 (1992).
- Lund, H., Hallam, J. & Lee, W. in *Proc. IEEE 3rd Int. Conf. on Evolutionary Computation* (eds Fukuda, T. et al.) 384–389 (IEEE Press, Piscataway, NJ, 1996).
- Thompson, A. in *Evolutionary Robotics: From Intelligent Robotics to Artificial Life* (ER'97) (ed. Gomi, T.) 101–125 (AAI Books, Ontario, 1997).
- Funes, P. & Pollack, J. Evolutionary body building: adaptive physical designs for robots. *Artif. Life* **4**, 337–357 (1998).
- Leger, C. *Automated Synthesis and Optimization of Robot Configurations: An Evolutionary Approach*. Thesis, Carnegie Mellon Univ. (1999).
- Kochan, A. Rapid prototyping trends. *Rapid Prototyp. J.* **3**, 150–152 (1997).
- Lenski, R. E., Ofria, C., Collier, T. & Adami, C. Genome complexity, robustness and genetic interactions in digital organisms. *Nature* **400**, 661–664 (1999).
- Joy, B. Why the future doesn't need us. *WIRED Magazine* **8**(4), 238–264 (2000).
- Watson, R. A., Ficici, S. G. & Pollack, J. B. in '99 *Congress on Evolutionary Computation* (eds Angeline, P. et al.) 335–342 (IEEE Press, New York, 1999).
- Beer, R. D. *Intelligence as Adaptive Behavior* (Academic, Boston, 1990).
- Ziemelis, K. Putting it on plastic. *Nature* **393**, 619–620 (1998).
- Baughman, R. H. et al. Carbon nanotube actuators. *Science* **284**, 1340–1344 (1999).
- Moravec, H. *Robot—From Mere Machine to Transcendent Mind* (Oxford Univ. Press, 1999).
- Mahfoud, S. W. *Niching Methods for Genetic Algorithms*. Thesis, Univ. Illinois at Urbana-Champaign (1995).
- Rechenberg, I. *Evolutionstrategie: Optimierung Technischer Systeme nach Prinzipien der Biologischen Evolution* (Frommann-Holzboog, Stuttgart, 1973).
- Fogel, L. J., Owens, A. J. & Walsh, M. J. *Artificial Intelligence through Simulated Evolution* (Wiley, New York, 1966).
- Holland, J. *Adaptation in Natural and Artificial Systems* (Univ. Michigan Press, Ann Arbor, 1975).
- Koza, J. *Genetic Programming* (MIT Press, Cambridge, MA, 1992).
- Gruau, F. & Quatramaran, K. in *Proc. 4th Eur. Conf. on Artificial Life* (eds Husbands, P. & Harvey, I.) (MIT Press, Cambridge, MA, 1997).
- Chellapilla, K. & Fogel, D. Evolution, neural networks, games and intelligence. *Proc. IEEE* **87**, 1471–1496 (1999).
- Hillis, D. in *Artificial Life II* (eds Langton, C., Taylor, J. F. & Rasmussen, S.) 313–322 (Addison-Wesley,

Reading, Massachusetts, 1992).

30. Pollack, J. B. & Blair, A. D. Co-evolution in the successful learning of backgammon strategy. *Machine Learning* **32**, 225–240 (1998).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or from the London editorial office of Nature.

Acknowledgements

We thank the DEMO Lab members for useful discussions: P. Funes, R. Watson, O. Melnik, S. Ficici, G. Hornby, E. Sklar & S. Levy. We also thank K. Quigley and G. Widberg for technical assistance. This research was partially supported by the Defense Advanced Research Projects Agency and by the Fischbach Postdoctoral Fellowship.

Correspondence and requests for materials should be addressed to the authors (e-mail: lipson@cs.brandeis.edu or pollack@cs.brandeis.edu).

Acclimation of ecosystem CO₂ exchange in the Alaskan Arctic in response to decadal climate warming

Walter C. Oechel*, George L. Vourlitis†*, Steven J. Hastings*, Rommel C. Zulueta*, Larry Hinzman‡ & Douglas Kane‡

* Global Change Research Group, Department of Biology,

San Diego State University, San Diego, California 92182-4614, USA

† Biological Sciences Program, California State University, San Marcos, California 92096-0001, USA

‡ Institute of Water Resources, University of Alaska, Fairbanks, Alaska 99706, USA

Long-term sequestration of carbon in Alaskan Arctic tundra ecosystems was reversed by warming and drying of the climate in the early 1980s, resulting in substantial losses of terrestrial carbon^{1,2}. But recent measurements suggest that continued warming and drying has resulted in diminished CO₂ efflux, and in some cases, summer CO₂ sink activity^{3,4}. Here we compile summer CO₂ flux data for two Arctic ecosystems from 1960 to the end of 1998. The results show that a return to summer sink activity has come during the warmest and driest period observed over the past four decades, and indicates a previously undemonstrated capacity for ecosystems to metabolically adjust to long-term (decadal or longer) changes in climate. The mechanisms involved are likely to include changes in nutrient cycling, physiological acclimation, and population and community reorganization. Nevertheless, despite the observed acclimation, the Arctic ecosystems studied are still annual net sources of CO₂ to the atmosphere of at least 40 g C m⁻² yr⁻¹, due to winter release of CO₂, implying that further climate change may still exacerbate CO₂ emissions from Arctic ecosystems.

Arctic ecosystems have historically been net sinks for atmospheric CO₂ due to predominance of cold, wet soils that effectively reduced rates of decomposition of organic matter⁵. However, warming and associated soil drying has recently stimulated rates of Arctic ecosystem (plant + soil) respiration relatively more than rates of gross primary production, resulting in a net loss of CO₂ from Alaskan^{1–2,6} and Siberian⁷ tundra ecosystems. Short-term (days–years) field and laboratory studies revealed that small fluctuations in soil water content, and to a lesser extent temperature, alter the net CO₂ flux of Arctic ecosystems^{8,9}. But long-term (years–decades) trends in Arctic ecosystem function will probably differ from those observed over the short term. For example, although climate change initially results in rapid rates of soil organic matter decomposition (and

net ecosystem carbon loss), such losses will probably decrease over time as pools of labile carbon decline¹⁰. Furthermore, initially rapid rates of soil organic matter decomposition should eventually lead to adjustments in nitrogen-mineralization rates, and possibly, shifts in soil and plant C:N ratios and plant species composition, which can stimulate long-term sequestration of carbon^{11–13}. Because many ecosystem processes (for example, mineralization of nitrogen and vegetation distribution) respond to environmental perturbations over longer timescales than do microbial respiration and leaf photosynthesis, the ecosystem response to climate change will vary over time as well¹⁴. The large (about 400–500 Gt C) carbon reserves in northern soils and permafrost⁵, and the importance of northern ecosystems in global climate patterns and processes¹⁵, highlight the need for understanding the effects of climate variability on long-term (decadal or longer) Arctic ecosystem function.

We analysed net CO₂ flux (Figs 1a and 2a, and Supplementary Information) and meteorological data (Figs 1b,c and 2b,c) collected over the past 39 years (1960–98) to assess the long-term trends in Alaskan Arctic ecosystem net CO₂ flux and climate change. Net CO₂ flux data were derived from a variety of field campaigns (see Supplementary Information), many of which involved the authors. Summer (June–August) mean air temperature and total precipitation data were obtained for several measurement stations ($n = 9$) in northern Alaska. These data were used to assess the change in ambient temperature and surface water balance (calculated as the

difference between precipitation and potential evapotranspiration: $P - PET$) of Alaskan coastal plain and interior regions over the past four decades.

Average summer air temperature increased by $0.05^{\circ}\text{C yr}^{-1}$ for both coastal plain (Fig. 1b) and interior regions (Fig. 2b) over the 39-year climate record. Although there was no significant temporal trend in precipitation (Figs 1c and 2c), the surface water balance ($P - PET$) decreased by 2.6 mm yr^{-1} for the coastal plain (Fig. 1d) and 6.7 mm yr^{-1} for interior regions (Fig. 2d). The increase in temperature and surface water deficit ($P - PET$) led to a change in the magnitude and direction of Arctic ecosystem net CO₂ flux. For example, wet sedge ecosystems were net CO₂ sinks of $-25 \text{ g C m}^{-2} \text{ yr}^{-1}$ between 1970 and 1972 (negative values denote net ecosystem CO₂ uptake; Fig. 1a), while moist-tussock tundra ecosystems were reported to be net sinks of $-120 \text{ g C m}^{-2} \text{ yr}^{-1}$ before 1976 (Fig. 2a). This sink activity coincided with a period when ambient temperature was on average $0.6\text{--}0.7^{\circ}\text{C}$ cooler than the overall (39-year) mean ambient temperature (Figs 1b and 2b) and the surface water balance was 30–50 mm higher than the overall average $P - PET$ (Figs 1d and 2d). After the mid-1980s and/or the early 1990s, however, wet-sedge tundra was a net source of CO₂ to the atmosphere of on average $25 \text{ g C m}^{-2} \text{ yr}^{-1}$ (Fig. 1a) while moist-tussock tundra ecosystems were losing between $50\text{--}450 \text{ g C m}^{-2} \text{ yr}^{-1}$ (Fig. 2a). This net CO₂ loss coincided with a period when the ambient temperature was 0.5°C higher than the overall mean temperature (Figs 1b and 2b) and $P - PET$ was 35 and 50–100 mm lower than the overall average coastal plain (Fig. 1d) and interior surface water balance (Fig. 2d), respectively.

In recent years, the net ecosystem CO₂ flux of Alaskan Arctic tundra ecosystems has acclimated to climate warming, and between 1992 and 1996, net summer source activity declined and eventually

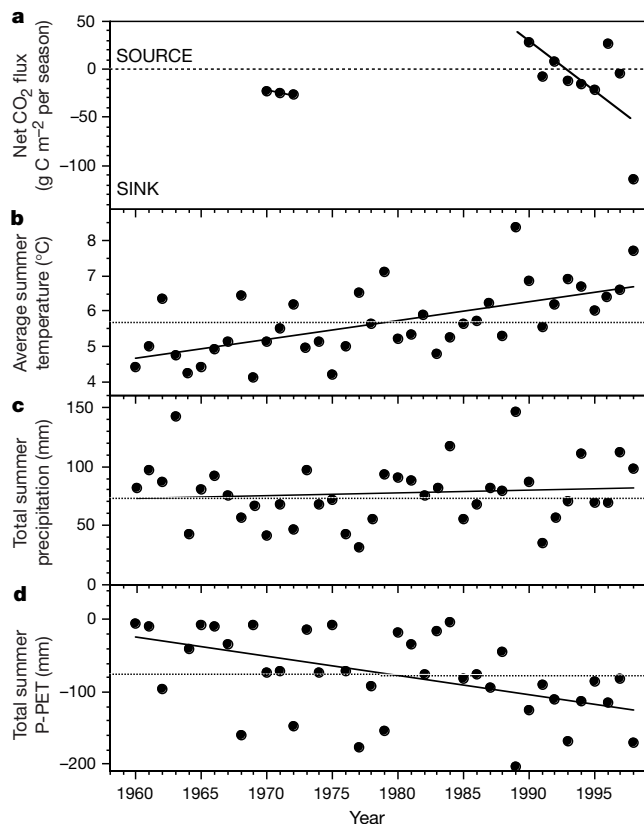


Figure 1 Long-term trends in summer net CO₂ flux and climate for Alaskan coastal wet-sedge tundra. **a**, Summer (June–August) net CO₂ exchange for coastal wet-sedge tundra ecosystems of the North Slope of Alaska. Net CO₂ exchange data before 1991 are derived from literature values, while data collected after 1991 are from chamber and eddy covariance measurements. Positive values depict net CO₂ loss to the atmosphere. Horizontal dashed lines indicate zero net flux. **b**, The average summer air temperature, **c**, total summer precipitation, and **d**, total summer precipitation minus potential evapotranspiration ($P - PET$) for Barrow, Kotzebue, Barter Island, Prudhoe Bay and Cape Lisburne between 1960 and 1998. Bold lines are based on linear regressions and horizontal dashed-lines indicate the 39-year average value.

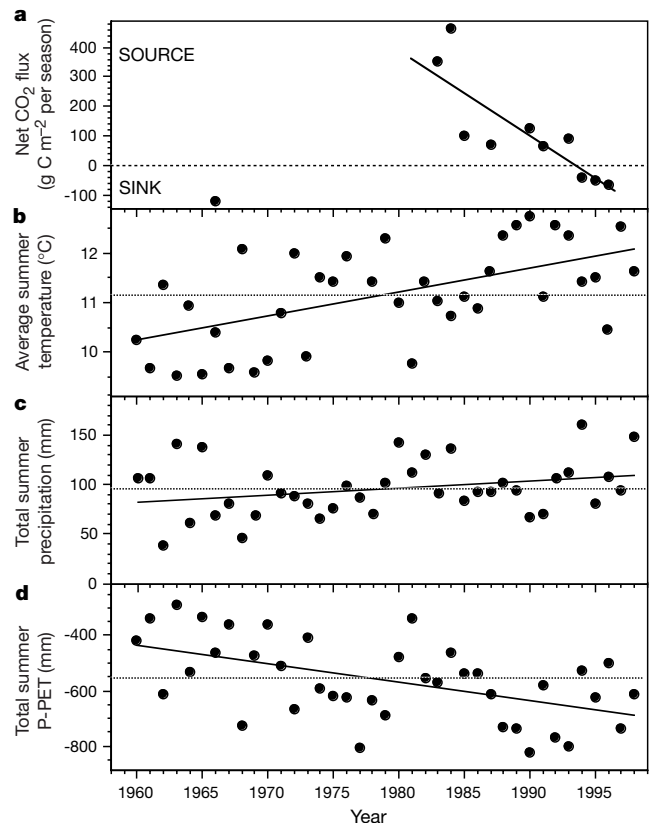


Figure 2 Long-term trends in summer net CO₂ flux and climate for Alaskan interior moist-tussock tundra. **a**, As Fig. 1 but for interior moist-tussock tundra ecosystems of the North Slope of Alaska. **b–d**, As for Fig. 1 but for Bettles, Arctic Village, Chandalar Lake and Umiat. Bold lines are based on linear regressions and horizontal dashed lines indicate the 39-year average value.

disappeared (Figs 1a and 2a, and Supplementary Information). The net summer CO_2 flux of wet-sedge tundra was in balance between 1991 and 1992, but after 1992, these ecosystems were net summer sinks of between -12 and -20 g C m^{-2} over the summer growing season (-0.12 and $-0.34 \text{ g C m}^{-2} \text{ d}^{-1}$) up to 1995 (Fig. 1a, and Supplementary Information). Similarly, moist-tussock tundra exhibited a seasonal net CO_2 loss of $100 \text{ g C m}^{-2} \text{ yr}^{-1}$ ($0.88 \text{ g C m}^{-2} \text{ d}^{-1}$) between 1990 and 1993, but after 1994, moist-tussock tundra was consistently accumulating on average -50 g C m^{-2} ($-0.54 \text{ g C m}^{-2} \text{ d}^{-1}$) during the summer growing season (Fig. 2a, and Supplementary Information). The increase in net CO_2 uptake (and/or decrease in net CO_2 efflux) observed after 1991 coincided with the highest average summer temperature (Figs 1b and 2b) and surface water deficit (Figs 1d and 2d) observed for the entire 39-year period.

Although recent measurements indicate that Arctic ecosystems are net sinks for atmospheric CO_2 over the summer growing season, net losses of CO_2 over annual timescales continue because of winter CO_2 losses^{7,16}. For example, wet-sedge tundra near Prudhoe Bay, Alaska, accumulated on average -5 g C m^{-2} during the summer growing seasons of 1994–97, but when combined with winter (September–May) losses of 44 g C m^{-2} measured between 1996 and 1997, the average annual net CO_2 efflux was 39 g C m^{-2} (Fig. 3).

No long-term observation program at standard locations has been in place over the past four decades, and thus, the long-term net CO_2 flux database consists of data collected from several sites and field observational studies. Potential significant measurement bias from enclosure and micrometeorological (that is, eddy covariance) sampling techniques were carefully evaluated and ruled out as a source of error¹⁷. Site-specific differences in ecosystem properties (for example, plant species composition, productivity, and hydrology) could also bias the long-term net CO_2 flux record. However, cross-calibration of research sites, and good correspondence between fluxes measured at different research sites in the same physiographic region (see Supplementary Information), suggest that potential errors associated with the utilization of data collected from a variety of research locations were minimal.

Although the mechanism responsible for the observed acclimation in Arctic ecosystem net CO_2 exchange is uncertain, climate change may initially result in large transient states¹⁸, such as the

large net CO_2 efflux observed during the mid-1980s and early 1990s. Simulation studies suggest that climate warming and concomitant changes in hydrology will ultimately lead to increased Arctic ecosystem CO_2 sequestration^{12,13,18–20}. However, the effects may not be fully manifest for decades and the initial change in ecosystem CO_2 flux may be opposite in direction to the long-term change^{12,20}.

The large net CO_2 loss observed during the mid-1980s and early 1990s implies relatively higher rates of organic matter decomposition, and ultimately, N-mineralization. However, simulation and empirical studies indicate that younger, more labile carbon is rapidly decomposed during the initial stages in soil warming and drying^{10,12}. If the N requirements of decomposers are not initially met by this labile litter and organic matter, N may be temporarily immobilized in microbial biomass²¹, and subsequent increases in net ecosystem CO_2 assimilation may lag behind initial increases in microbial respiration^{12,13,19,20}. Because the growth of Arctic plants is strongly N-limited²², enhanced rates of net N-mineralization should eventually stimulate rates of gross primary production and atmospheric CO_2 sequestration^{12,13,19}. Thus, rapid initial rates of net CO_2 efflux caused by climate warming and drying could be balanced over time by N-induced increases in gross primary production, which is consistent with the long-term net CO_2 flux record.

A shift in plant species composition towards more productive functional groups would further dampen transient CO_2 losses over time. There is evidence that the relative abundance of deciduous shrubs has increased in response to climate change over the past 1–2 decades in Alaskan moist-tussock tundra ecosystems¹¹. This change in community composition has significant implications for net CO_2 exchange because deciduous shrubs have intrinsically more rapid rates of N-uptake, productivity, and maximum photosynthesis²³. Because changes in higher plant species composition occur over decade timescales, the effects of community-level changes should result in a gradual increase in ecosystem productivity. This hypothesis is consistent with the recent trends in net CO_2 exchange (Figs 1a and 2a) and the observed temporal trends in the high-latitude Noenalized Difference Vegetation Index (NDVI)²⁴.

The results from this study should not be taken to imply that high-latitude warming and alterations in hydrology have only short-term effects on Arctic ecosystem function and net CO_2 flux. The large net summer efflux of the 1980s and early 1990s followed a period of rapid warming during the mid-1970s. Presumably, another period of rapid warming would result in another increase in growing-season net CO_2 emissions, which would be expected to exacerbate further the current annual loss of CO_2 to the atmosphere. Continuous long-term studies of interannual variability and change in climate, ecosystem and ecosystem function will enhance our ability to predict the multi-decadal effects of climate change on Arctic ecosystems and carbon flux. □

Methods

CO_2 flux measurement

Net CO_2 flux data were compiled from literature surveys and field measurements. Data for 1966 are from chamber measurements conducted near Meade River, Alaska²⁵. Data for 1970–72 are from aerodynamic, chamber, and simulation studies conducted during the International Biological Program (IBP) research of moist/wet-sedge ecosystems near Barrow, Alaska^{26,27}. Data from 1983–85 and 1987 were collected from null-balance, environmentally controlled chambers in moist-tussock tundra ecosystems at Toolik Lake, Alaska^{1,6}. Data from moist-tussock and moist/wet-sedge tundra from 1990–92 are from dynamic, closed-chamber measurements¹. Sources and methods of measurement for net CO_2 flux data collected after 1992 are from a variety of methods (see Supplementary Information^{28,29}).

Chamber flux measurements of net CO_2 flux³⁰ were made on 5–12, 0.5-m² plots per site between 1 June and 31 August. Net CO_2 flux was calculated as the time rate of change in CO_2 concentration in the chamber atmosphere measured over a 60–180 s measurement interval. Net CO_2 flux was measured every 1–1.5 h over a 24-h cycle, and repeated every 7 d throughout the growing season. Eddy covariance measurements^{3,4} were conducted at either 2.5 m or 3.5 m above ground level. Vertical wind speed and temperature fluctuations were measured at 10 Hz using a three-dimensional sonic anemometer-thermometer, while CO_2 and H_2O vapour fluctuations were measured using fast-response (10 Hz) closed- and open-path infrared gas analysers. Raw CO_2 and H_2O vapour fluctuations were output as

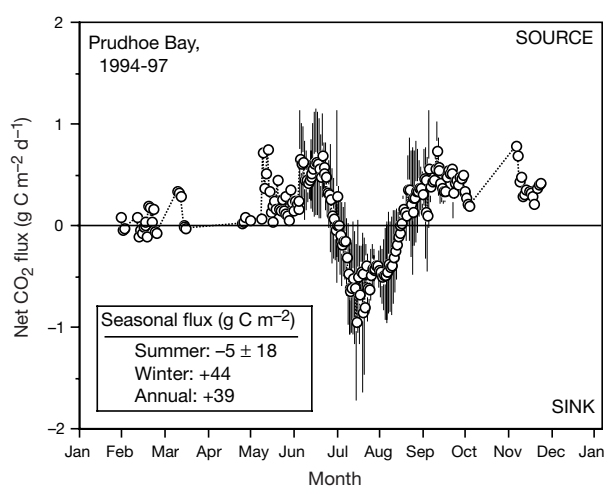


Figure 3 Daily integrated net CO_2 flux for a moist/wet sedge coastal tundra ecosystem near Prudhoe Bay ($70^\circ 16' \text{ N}$, $148^\circ 53' \text{ W}$). Net CO_2 flux was measured using eddy covariance. Data points for summer months (June–August) represent a mean (error bars, $\pm 1 \text{ s.d.}$, $n = 4$ years) calculated from measurements made between 1994 AND 97, while data for winter months (September–May) are from measurements made during the 1996–97 winter period. Values in the inset correspond to the integrated net CO_2 exchange calculated for the summer, winter and annual periods. Positive values depict net CO_2 loss to the atmosphere.

mean voltages and converted to concentrations by multiplying by the requisite calibration constant. Net CO₂ flux was computed following coordinate rotation, and was calculated and stored on a personal computer as 30-min averages using a 200-s running mean and digital recursive filtering technique. Closed-path CO₂ flux was corrected for the simultaneous flux of H₂O vapour, while open-path CO₂ flux was corrected for the simultaneous fluctuations in both heat and H₂O vapour. Spectral analyses and energy balance closure (between 3% and 13%) indicated good system performance during the summer measurement periods^{3,4}.

Cross-comparisons of techniques used to measure net ecosystem CO₂ fluxes, including chamber, eddy covariance, and aerodynamic measurements indicate good agreement and no systematic bias in the trends due to the methodology employed¹⁷ (see Supplementary Information). For example, flux data derived from paired chamber and aerodynamic studies conducted during the IBP (1970–72) were shown to vary by only 9 g C m⁻² yr⁻¹ (between -32 and -40.7 g C m⁻² yr⁻¹) (L. L. Tieszen, unpublished data). Cross-calibration of net CO₂ flux data derived from different measurement sites indicates that site-specific bias in net CO₂ flux was negligible (see Supplementary Information). For example, net CO₂ flux data from Happy Valley and Toolik Lake during the 1990–91 growing seasons were nearly identical (mean ± 1 s.d. daily net CO₂ flux averaged over 1990–91 = 1.01 ± 0.47 and 1.02 ± 0.20 g C m⁻² d⁻¹ for Toolik Lake and Happy Valley, respectively; see Supplementary Information). Differences in net CO₂ exchange measured over a partial growing season (mid-June to late-July) in wet-sedge ecosystems of Barrow and Prudhoe Bay in 1993 were also negligible (-0.45 g C m⁻² d⁻¹ for Barrow and -0.46 g C m⁻² d⁻¹ for Prudhoe Bay; W.C.O., unpublished data).

Climate measurement

Monthly temperature and precipitation data were compiled from the Desert Research Institute (DRI), Western Region Climate Center data archives (<http://www.wrcc.dri.edu/summary/climsmak.html>). Alaskan coastal-plain data are from Barrow, Kotzebue, Barter Island, Prudhoe Bay and Cape Lisburne, while Alaskan interior data are from Bettles, Arctic Village, Chandalar Lake and Umiat. Missing data for coastal-plain and interior sites were estimated using linear regression from the Barrow and Bettles data sets, respectively, which were the most complete databases for the entire 39-year period. Mean summer (June–August) temperature was calculated from average daily values, and total summer precipitation was calculated from total daily precipitation data. Potential evapotranspiration (PET) was calculated as a function of the mean monthly temperature³, and the surface moisture index (P – PET) was calculated as the difference between total monthly precipitation and total monthly PET. Summer estimates of P – PET were summed from the monthly estimates of P – PET.

Received 29 September 1999; accepted 4 July 2000.

- Oechel, W. C. *et al.* Recent change of arctic tundra ecosystems from a carbon sink to a source. *Nature* **361**, 520–523 (1993).
- Oechel, W. C., Vourlitis, G. L., Hastings, S. J. & Bochkarev, S. A. Change in carbon dioxide flux at the U.S. Tundra International Biological Program Sites at Barrow, AK. *Ecol. Appl.* **5**, 846–855 (1995).
- Vourlitis, G. L. & Oechel, W. C. Landscape-scale CO₂, H₂O vapor, and energy flux of moist-wet coastal tundra ecosystems over two growing seasons. *J. Ecol.* **85**, 575–590 (1997).
- Vourlitis, G. L. & Oechel, W. C. Eddy covariance measurements of net CO₂ flux and energy balance of an Alaskan moist-tussock tundra ecosystem. *Ecology* **80**, 686–701 (1999).
- Miller, P. C., Kendall, R. & Oechel, W. C. Simulating carbon accumulation in northern ecosystems. *Simulation* **40**, 119–131 (1983).
- Grulke, N. E. *et al.* Carbon balance in tussock tundra under ambient and elevated atmospheric CO₂. *Oecologia* **83**, 485–494 (1990).
- Zimov, S. A. *et al.* Siberian CO₂ efflux in winter as a CO₂ source and cause of seasonality in atmospheric CO₂. *Clim. Change* **33**, 111–120 (1996).
- Billings, W. D., Luken, J. O., Mortensen, D. A. & Peterson, K. M. Arctic tundra: A source or sink for atmospheric carbon dioxide in a changing environment? *Oecologia* **52**, 7–11 (1982).
- Oechel, W. C., Vourlitis, G. L., Hastings, S. J., Ault, R. P. & Bryant, P. The effects of water table manipulation and elevated temperature on the net CO₂ flux of wet sedge tundra ecosystems. *Glob. Change Biol.* **4**, 77–90 (1998).
- Christensen, T. R., Jonasson, S., Callaghan, T. V. & Havström, M. On the potential CO₂ releases from tundra soils in a changing climate. *Appl. Soil Ecol.* **11**, 127–134 (1999).
- Chapin, F. S. III, Shaver, G. R., Giblin, A. E., Nadelhoffer, K. J. & Laundre, J. A. Responses of arctic tundra to experimental and observed changes in climate. *Ecology* **76**, 694–711 (1995).
- Rastetter, E. B. *et al.* in *Global Change and Arctic Terrestrial Ecosystems* (eds Oechel, W. C. *et al.*) 437–451 (Springer, New York, 1997).
- McKane, R. B. *et al.* Reconstruction and analysis of historical changes in carbon storage in arctic tundra. *Ecology* **78**, 1188–1198 (1997).
- Oechel, W. C. & Vourlitis, G. L. The effects of climate change on arctic tundra ecosystems. *Trends Ecol. Evol.* **9**, 324–329 (1994).
- Bonan, G. B., Pollard, D. & Thompson, S. L. Effects of boreal forest vegetation on global climate. *Nature* **359**, 716–718 (1992).
- Oechel, W. C., Vourlitis, G. L. & Hastings, S. J. Cold-season CO₂ emission from arctic soils. *Glob. Biogeochem. Cycles* **11**, 163–172 (1997).
- Oechel, W. C., Vourlitis, G. L., Brooks, S., Crawford, T. L. & Dumas, E. Intercomparison among chamber, tower, and aircraft net CO₂ and energy fluxes measured during the Arctic Systems Science Land-Atmosphere-Ice-Interactions (ARCSS-LAI) flux study. *J. Geophys. Res.* **103**, 28993–29004 (1998).
- Smith, T. M. & Shugart, H. H. The transient response of terrestrial carbon storage to a perturbed climate. *Nature* **361**, 523–526 (1993).
- McGuire, A. D. *et al.* Interactions between carbon and nitrogen dynamics in estimating net primary productivity for potential vegetation in North America. *Glob. Biogeochem. Cycles* **6**, 101–124 (1992).
- Waelbroeck, C., Mofray, P., Oechel, W. C., Hastings, S. & Vourlitis, G. The impact of permafrost thawing on the carbon dynamics of tundra. *Geophys. Res. Lett.* **24**, 229–232 (1997).

- Morehead, D. L. & Reynolds, J. F. Effects of climate change on decomposition in arctic tussock tundra: A modeling synthesis. *Arct. Alp. Res.* **25**, 403–412 (1993).
- Chapin, F. S. III & Shaver, G. R. Individualistic growth response of tundra plant species to environmental manipulations in the field. *Ecology* **66**, 564–576 (1985).
- Oberbauer, S. F. & Oechel, W. C. Maximum CO₂ assimilation rates of vascular plants on an Alaskan arctic tundra slope. *Hol. Ecol.* **12**, 312–316 (1989).
- Myneni, R. B., Keeling, C. D., Tucker, C. J., Asrar, G. & Nemani, R. R. Increasing plant growth in the northern high latitudes from 1981 to 1991. *Nature* **386**, 698–702 (1997).
- Johnson, P. L. & Kelley, J. J. Dynamics of carbon dioxide and productivity in an arctic biosphere. *Ecology* **51**, 73–80 (1970).
- Coyne, P. I. & Kelley, J. J. CO₂ exchange over the Alaskan arctic tundra: meteorological assessment by an aerodynamic method. *J. Appl. Ecol.* **12**, 587–611 (1975).
- Tieszen, L. L. CO₂ exchange in the Alaskan arctic tundra: Seasonal changes in the rate of photosynthesis of four species. *Photosynthetica* **9**, 376–390 (1975).
- Oechel, W. C. & Vourlitis, G. L. in *Advances in Soil Science: Soils and Global Change* (eds Lal, R., Kimbel, J., Levine, E. & Stewart, B. A.) 117–129 (Lewis, Boca Raton, 1995).
- Vourlitis, G. L. *et al.* Physiological models for scaling plot measurements of CO₂ flux across an arctic tundra landscape. *Ecol. Appl.* **10**, 60–72 (2000).
- Vourlitis, G. L., Oechel, W. C., Hastings, S. J. & Jenkins, M. A. A system for measuring *in situ* CO₂ and CH₄ flux in unmanaged ecosystems: An arctic example. *Funct. Ecol.* **7**, 369–379 (1993).
- Thornthwaite, C. J. An approach toward rational classification of climate. *Geogr. Rev.* **38**, 1–94 (1948).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank the Alaska Department of Natural Resources, US Bureau of Land Management, North Slope Borough Department of Industrial Development, ARCO Alaska, Inc., and BP Exploration for access to the North Slope study sites. Field assistance from R. Ault, P. Bryant and M. Vourlitis, and technical expertise from T. Crawford, T. Meyers, R. McMillen and Y. Harazono are acknowledged. We also thank A. D. McGuire, T. Christensen and S. Verma for commenting on earlier versions of the manuscript. This work was supported by the National Science Foundation, Arctic Systems Science, Land-Atmosphere-Ice-Interactions Program and Division of Environmental Biology, Joint Program on Terrestrial Ecology and Global Change, and by the US Department of Energy, Office of Health and Environmental Research. Logistic support was provided by the Northern Arctic Research Laboratory, the Polar Ice Coring Office (PICO), University of Alaska, Fairbanks (1994), PICO-University of Nebraska, Lincoln (1995–97), Institute of Arctic Biology, University of Alaska, Fairbanks, Piquini Management Corporation and the North Slope Borough Department of Wildlife Management.

Correspondence and requests for materials should be addressed to W.C.O. (e-mail: oechel@sunstroke.sdsu.edu).

Diagenetic origin of quartz silt in mudstones and implications for silica cycling

Jürgen Schieber*, Dave Krinsley*† & Lee Ricuputi‡

* Department of Geology, University of Texas at Arlington, Arlington, Texas 76019, USA

† Department of Geological Sciences, University of Oregon, Eugene, Oregon 97403, USA

‡ Chemical and Analytical Sciences Division, Oak Ridge National Laboratory, PO Box 2008, Oak Ridge, Tennessee 37831-6365, USA

Mudstone—the most abundant sedimentary rock type¹, composed primarily of clay- or silt-sized particles—contains most of the quartz found in sedimentary rocks². These quartz grains, which are chemically and mechanically resistant and therefore preserve their characteristics well, have long been considered to be derived from the continental crust¹. Here we analyse quartz silt from black shales in the eastern USA, dating back to the Late Devonian period (about 370 million years ago), using backscattered electron and cathodoluminescence imaging and measure oxygen isotopes with an ion probe. Our results indicate that up to 100% of the quartz silt in our samples does not originate from the continental crust. Instead, it appears to have precipitated early in diagenesis in algal cysts and other pore spaces³, with silica derived

from the dissolution of opaline skeletons of planktonic organisms, such as radiolaria and diatoms. Transformation of early diatoms into *in situ* quartz silt might explain the time gap between the earliest fossil occurrences of diatoms about 120 Myr ago⁴ and molecular evidence for a much earlier appearance between 266 or even 500 Myr ago^{5,6}. Moreover, if many other mudstone successions show similarly high proportions of *in situ* precipitated—rather than detrital—quartz silt, the sedimentary record in mudstones may have been misinterpreted in the past, with consequences for our estimates of palaeoproductivity as well as our perceptions of the dynamics and magnitude of global biogeochemical cycling of silica.

We examined quartz silt grains (4–62 μm) in mudstones from the Late Devonian Chattanooga and New Albany Shales of the eastern USA with scanning cathodoluminescence (using a scanning electron microscope; SEM-CL), because recent studies have revealed textures that can be used to determine quartz provenance even in very small grains^{7,8}. Grey-scale pictures were produced using mono-CL rather than colour-CL, as the former is much faster at producing images with textural details at a spatial resolution of about 1 μm , and for this study textural features are far more critical than colour or CL wavelength. Two principal types of quartz silt were differentiated (Fig. 1): (1) grains with uniform to mottled luminescence (Fig. 1b), and (2) grains that were essentially non-luminescent (Fig. 1b). Whereas grains of the first type compare well to metamorphic quartz described in a prior SEM-CL study⁷, and were expected because of abundant metamorphic source rocks in

the Acadian orogen to the east⁹, the source of the essentially non-luminescent grains of the second type is more problematic. Quartz of this type has not been observed previously by SEM-CL⁷, but in light-microscope CL studies low-luminescence quartz has commonly been interpreted as being of low-temperature authigenic/diagenetic origin^{10,11}.

To verify this, we examined by SEM-CL quartz that had been identified as authigenic by other methods in previous studies of these rocks^{3,12}. In partially quartz cemented siltstone beds¹², metamorphic quartz is always considerably lighter than the dark authigenic quartz, although careful examination of this authigenic quartz reveals subtle CL variability and fine structure (Fig. 2b). Very early diagenetic quartz also fills large, sand-size (0.1–1.0 mm) algal cysts in these rocks³. These cysts, remains of planktonic green algae, have walls that consist of complex lipoid-like substances that are highly resistant to chemical breakdown and bacterial degradation⁴. Dissolving radiolarian tests, still preserved in carbonate and phosphate concretions within these mudstones³, provided silica for initial chalcedony precipitation in these cysts and associated pore spaces. Later on, these chalcedony grains recrystallized to single-crystal quartz grains³. Large, silica-filled cysts represent a source of *in situ* quartz sand and occur concentrated in lag deposits on erosion surfaces and sequence boundaries¹². Thus, even though these sand-size authigenic quartz grains constitute less than 1% of the deposited sediment volume, they are important for reconstructing the depositional history of these mudstones³. Although largely non-luminescent (Fig. 3a), these quartz sand grains also show fine

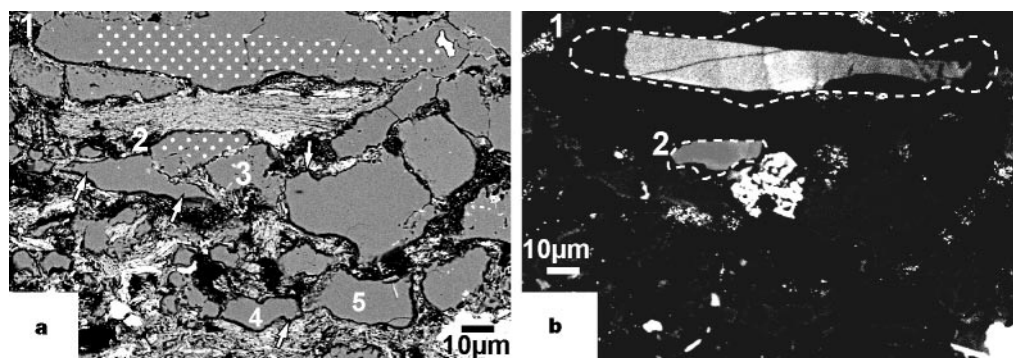


Figure 1 Quartz grains in a silt layer from Devonian mudstone succession. **a**, Backscattered electron (BSE) image (quartz grains are medium grey); **b**, SEM-CL image. We note in **a** embayments and lobate-to-pointed projections on a number of grains (white arrows), a first indication that they formed as diagenetic cyst fills. In contrast, quartz silt grains of detrital origin experience little rounding during transport, and typically have angular outlines (see also Fig. 2b) with sharp corners¹. Panel **b** shows that there are only three grains that have strong luminescence. Stippled areas in **a** mark the positions of presumed

metamorphic grains numbered 1 and 2. Dashed outlines in **b** mark BSE outlines of quartz grains with metamorphic cores. The cores of grains 1 and 2 have SEM-CL characteristics typically found in metamorphic quartz⁴, and grain 3 appears to be a metamorphic rock fragment with intergrown quartz, mica and feldspar. Grains 1 and 2 have noticeable overgrowth of diagenetic quartz. About 80% of the quartz visible in **a** is essentially non-luminescent, and presumed to be of diagenetic origin. The grains marked as number 4 and 5 in **a** are shown enlarged and enhanced in Fig. 4c and d, respectively.

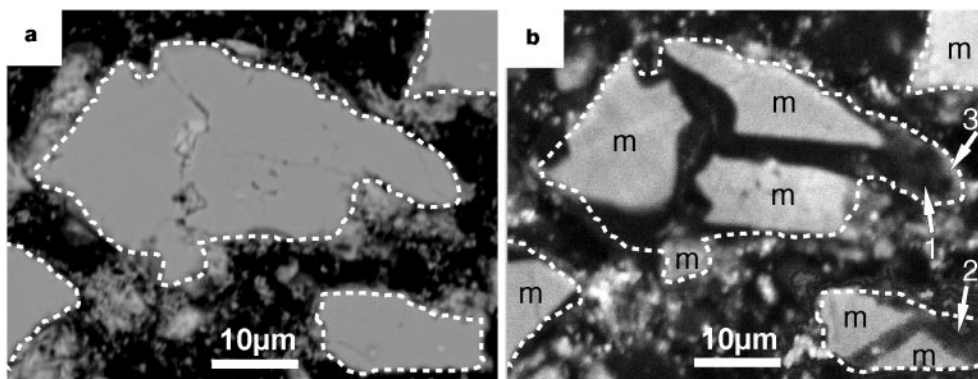


Figure 2 Detrital quartz silt with pore-filling diagenetic quartz. **a**, BSE image where quartz grains (outlined with dashed white line) are essentially uniform medium grey. **b**, SEM-CL image showing how metamorphic quartz grains (marked m) with light grey, uniform to mottled luminescence, have been fused together and overgrown by authigenic quartz of

black to dark grey colour (essentially non-luminescent). Arrows 1 and 2 show how quartz deposition not only occurred in the space between detrital grains, but also in the adjacent pore space. Arrow 3 points to a lighter-coloured rim (zonation) within pore-space-filling quartz.

structure through zones of variable luminescence (Fig. 3b). Other features that distinguish them from detrital sand grains are embayments and lobate-to-pointed projections³ that resulted from deformation of the cyst walls during early burial.

Extrapolating these observations from the rare sand-size quartz grains to the much more abundant silt-size quartz grains in these rocks (Figs 2, 3) strongly suggests that the low- to non-luminescent quartz silt grains seen in Fig. 1b are of authigenic origin as well.

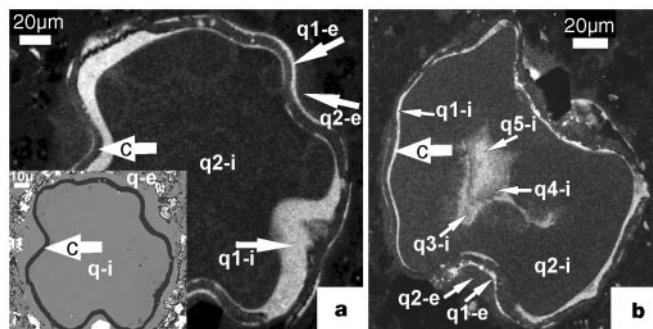


Figure 3 Several generations of diagenetic quartz growth in algal cysts. **a**, SEM-CL image of sand-size algal cyst that is filled and overgrown with quartz. The cyst underwent some deformation after deposition, and early internal silica deposition prevented collapse upon burial. Inset, BSE image showing the cyst wall (arrow c), quartz infill (q-i), and overgrowth (q-e). Although most of the diagenetic quartz is non-luminescent (q2-i and q2-e), there are thin rims of luminescent first-generation quartz (q1-i and q1-e). Panel **b** is similar to **a**, but shows additional growth stages (q3-i, q4-i and q5-i) with variable intensity of luminescence on the inside of the cyst (cyst wall marked with arrow c). Thus, although typically we see only a very thin or absent first generation (q1, moderately luminescent), and a very prominent second generation (q2, non-luminescent), of diagenetic quartz, some cysts may show a larger number of concentrically growing quartz generations. During reworking the organic cyst wall and outer quartz deposits (q-e) are removed³, and fragments of the latter are also a source of *in situ* quartz silt.

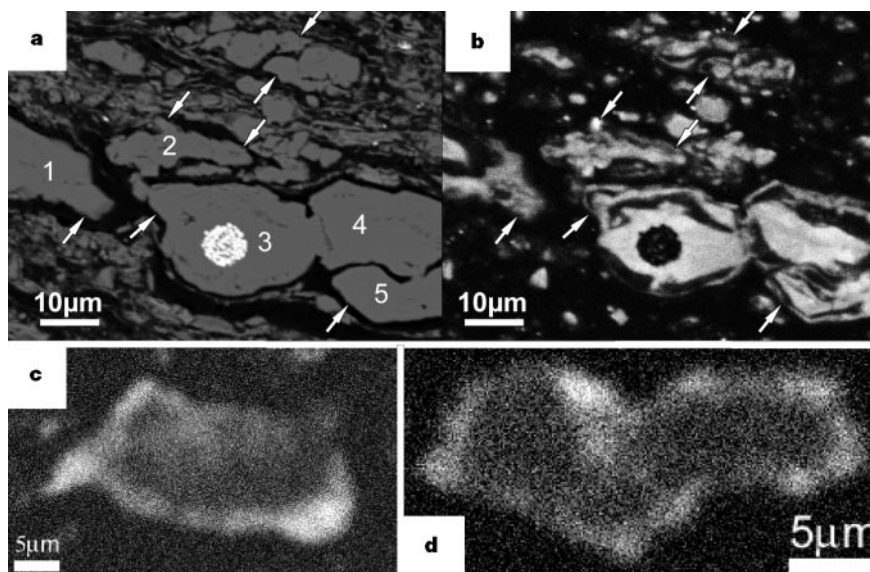


Figure 4 Quartz silt grains in mudstone matrix with several growth generations. **a**, **b**, A range of quartz grains in mudstone matrix. **a**, BSE image; quartz grains are medium grey with essentially uniform surface. **b**, SEM-CL image of the same region. Arrows in **a** point out embayments and lobate-to-pointed projections that are an indication that these quartz grains formed as diagenetic cyst fills and are *in situ* quartz grains³. Arrows from **a** are repeated in **b** for reference. Larger grains that are marked 1 to 5 in **a** show zoned luminescence patterns in SEM-CL view, suggesting that they are of *in situ* origin (compare Fig. 3b). Smaller grains in the upper half of the image also show zoned luminescence patterns. Quartz silt in this sample has highly positive $\delta^{18}\text{O}_{\text{VSMOW}}$ values and appears to be

Textural features (Fig. 4), such as zones of variable luminescence, embayments, and lobate projections, indicate that low-luminescence quartz silt probably also originated as early diagenetic infills of small algal cysts. The observation, in places, of thin, amber-coloured films (the cyst walls) surrounding this type of silt grain (best seen in tapered wedges at the edge of thin sections) supports this assumption. In the studied samples, 50% to almost 100% (for example, Fig. 4a, b) of the visible quartz silt shows the textural characteristics noted above, and we presume that it is of *in situ* authigenic origin. With an average quartz silt content of 40% (refs 12, 13), the authigenic quartz silt content of these Devonian mudstones may range from 20% to 40%. Although algal cysts were apparently the preferred locus of early diagenetic silica deposition in the case of our Devonian mudstones³, this need not be the case elsewhere. In the absence of algal cysts, interparticle voids and other fossil pores (for example, foraminifera tests) are alternative sites for early diagenetic silica deposition. As a consequence, some identifying textural features (for example, grain outline) may differ as well.

Because of the small grain size and limitations of instrument resolution, textural identification of silt-size authigenic quartz cannot be performed at the same level of confidence as is possible for larger, sand size grains. Nonetheless, our textural and CL discrimination of *in situ* authigenic quartz silt is confirmed by oxygen isotope measurements on individual quartz grains. Quartz silt grains texturally identified as either metamorphic or early diagenetic were analysed with a modified Cameca 4f ion microprobe at Oak Ridge National Laboratory using extreme energy filtering¹⁴. Spot size was about 15 μm , located in the centre of the grain, and precision was $\pm 1\text{‰}$ (1σ). Grains texturally identified as metamorphic have $\delta^{18}\text{O}_{\text{VSMOW}}$ values of $+9.4 \pm 1.5\text{‰}$ (1σ , $n = 7$), consistent with metamorphic quartz. Grains identified as *in situ* authigenic quartz have much higher $\delta^{18}\text{O}_{\text{VSMOW}}$ values of $+28.4 \pm 1.0\text{‰}$ (1σ , $n = 5$), typical for cherts and other types of diagenetic quartz¹⁵. Thus, textural identification of *in situ* quartz silt in these mudstones (Fig. 4) is confirmed by isotope data. That a

almost entirely of *in situ* origin. Grain 3 contains a pyrite framboid in the centre, further evidence for early diagenetic quartz deposition. Panels **c** and **d** are enlarged and enhanced images of quartz grains marked 4 (**d**) and 5 (**c**) in Fig. 1. They show internal zonation and an outer zone of brighter luminescent quartz, as seen in *in situ* quartz grains shown in Fig. 3. Thus, even though these quartz grains appear non-luminescent and featureless when contrasted with more luminescent metamorphic quartz (Fig. 1), they show the same internal features as found in *bona fide in situ* quartz grains. $\delta^{18}\text{O}_{\text{VSMOW}} = [({}^{18}\text{O}/{}^{16}\text{O})_{\text{sample}} / ({}^{18}\text{O}/{}^{16}\text{O})_{\text{VSMOW}}] - 1$, where VSMOW indicates Vienna standard mean ocean water.

large proportion of quartz silt in mudstones may not be detrital, but rather of *in situ* authigenic origin, has significant implications for a variety of research areas, including basin modelling, palaeoclimate, palaeoproductivity, and biogeochemical silica cycling.

It was long thought that the quartz silt content of mudstones diminishes in the offshore direction, and that the distribution of quartz silt in the mudstones of a sedimentary basin can be used to determine palaeocurrents¹ and distance to the shoreline¹⁶. If common, *in situ* quartz silt production could completely obscure expected patterns and lead to erroneous conclusions. For example, our Devonian mudstones contain 20–50% quartz silt in nearshore settings¹⁷, yet quartz silt contents 300–600 km offshore still average 40% (refs 12, 13), clearly at odds with the expected offshore decline of quartz silt abundance¹⁵. Our finding that a large portion of the quartz silt can be assigned to authigenic, *in situ*, silica deposition (Fig. 1) resolves this problem.

Previously, when large amounts of quartz silt were detected in distal mudstones, a high degree of sorting in the silt fraction was used to support aeolian transport as an explanation for such anomalies¹⁸. However, *in situ* quartz silt as described here forms in organic particles (for example, algal cysts) that have a limited size range, and thus is inherently well sorted. Because aeolian dust in mudstones is used as a palaeoclimate proxy¹⁹, mistaken identification of *in situ* quartz silt as aeolian will produce errors in palaeoclimate assessments.

The bulk of opaline silica produced by radiolaria and diatoms in ocean surface waters never becomes incorporated into deep-sea sediments because it dissolves while settling through several kilometres of water²⁰. In shallow epicontinental seas this dissolution effect is much less important, and a substantial portion of surface production can be buried. Although opal invariably dissolves in the sediment, when silica concentrations are large enough, authigenic quartz will precipitate²¹. Thus, *in situ* quartz silt can be considered a testament of vanished opaline constituents, and possibly as a proxy for surface productivity. If quantified and compared to organic-carbon and phosphorus contents, recognition of abundant *in situ* quartz silt in a carbonaceous mudstone could provide evidence for productivity control, and could contribute to the continuing discussion of whether carbonaceous mudstones owe their origin to high surface productivity or to enhanced preservation in anoxic basins^{22,23}. It might also help identify palaeoproductivity in sediments that are not characterized by abundant organic-carbon or phosphorus contents. The Devonian mudstones we studied are carbonaceous and were long considered a consequence of anoxia²⁴. Recent observations in support of shallow-water conditions and some oxygen in the water column^{12,17} suggest instead high surface productivity as a cause of abundant organic matter accumulation. This view is supported by our observation of large amounts of *in situ* quartz silt.

Identification of abundant authigenic quartz silt in mudstones may help explain the large gap in the geological record between the earliest fossil occurrences of diatoms (early Cretaceous period, about 120 Myr ago)⁴, and molecular evidence suggesting that they appeared during the Permian period (about 266 Myr ago)⁵ or the early Palaeozoic era⁶. This gap has been attributed to diatom dissolution in the past²⁵, and our results suggest that vanished diatoms may have been preserved as *in situ* quartz silt.

Failure to recognize this authigenic component may also affect our understanding of the dynamics and magnitude of the global biogeochemical cycling of silica¹⁹ and secular trends of silica cycling²⁶. As described here, *in situ* quartz silt is linked to opaline silica flux (radiolarians, diatoms), and thus to dissolved silica flux from continents and mid-oceanic ridges²⁰. Secular changes of the *in situ* quartz silt component in mudstones may therefore reflect changes in weathering and climate (continental dissolved flux), as well as changes in the rate of sea-floor spreading²⁰.

By volume, most mudstones are deposited along basin margins¹.

In these areas terrigenous clastic deposition dominates, and biogenic silica—the source of *in situ* quartz silt—will be of little consequence for the total sediment budget. Under those circumstances, quartz silt abundance is probably a useful indicator of shoreline proximity¹⁶. Distal mudstones, although of lesser volume, tend to cover large areas of depositional basins, occupy a disproportionate share of geological time, and contain a more complete sedimentary record. Owing to their slow accumulation, biogenic sedimentation is much more significant, and the *in situ* quartz silt component (as illustrated with our Late Devonian mudstones) should be most noticeable. Late Devonian carbonaceous mudstones occur over large portions of the North American continent²⁷, and mineralogical data from a broad range of localities show anomalous quartz contents²⁸. A study of the Late Devonian Antrim shale of the Michigan basin even reports that more than half of the quartz silt was of authigenic origin²⁹. Petrographic re-examination (by light microscope) of Late Devonian mudstones from across North America reveals abundant quartz silt grains that are texturally identical to those identified here as authigenic quartz infills of algal cysts.

So it seems that *in situ* authigenic quartz silt is an abundant and widespread component of Late Devonian carbonaceous mudstones in North America. Considering that many transgressive episodes in the Earth's history were accompanied by slowly deposited and often carbonaceous mudstone blankets³⁰, it is likely that many other mudstone successions contain a significant *in situ* quartz silt component. Although it remains to be determined how abundant this component is in mudstones of varying ages, conditions that produced our Late Devonian mudstones were not unique to that period in Earth history. If authigenic/intrabasinal quartz silt is widespread, a large portion of the sedimentary record may have been misinterpreted, with important implications in a variety of research areas. Although the identification of authigenic quartz silt is not a trivial matter, its recognition offers a variety of new opportunities to understand better the geological record. □

Received 15 July 1999; accepted 15 June 2000.

- Potter, P. E., Maynard, J. B. & Pryor, W. A. *Sedimentology of Shale* (Springer, New York, 1980).
- Blatt, H. Provenance studies and mudrocks. *J. Sedim. Petrol.* **55**, 69–75 (1985).
- Schieber, J. Early diagenetic silica deposition in algal cysts and spores: A source of sand in black shales? *J. Sedim. Res.* **66**, 175–183 (1996).
- Tappan, H. *The Paleobiology of Plant Protists* (Freeman, San Francisco, 1980).
- Kooistra, W. H. C. F. & Medlin, L. K. Evolution of the diatoms (Bacillariophyta): IV. A reconstruction of their age from small subunit rRNA coding regions and the fossil record. *Mol. Phylogenet. Evol.* **6**, 391–407 (1996).
- Philippe, H. et al. Comparison of molecular and paleontological data in diatoms suggests a major gap in the fossil record. *J. Evol. Biol.* **7**, 247–265 (1994).
- Seyedolali, A. et al. Provenance interpretation of quartz by scanning electron microscope-cathodoluminescence fabric analysis. *Geology* **25**, 787–790 (1997).
- Milliken, K. L. Cathodoluminescent textures and the origin of quartz silt in Oligocene mudrocks, south Texas. *J. Sedim. Res.* **64**, 567–571 (1994).
- Woodrow, D. L. in *The Catskill Delta* (eds Woodrow, D. L. & Sevon, W. D.) 51–63 (Special Paper 201, Geological Society of America, Boulder, Colorado, 1985).
- Matter, A. & Ramseyer, K. in *Provenance of Arenites* (ed. Zuffa, G. G.) 191–211 (Reidel, Dordrecht, 1985).
- Zinkernagel, U. Cathodoluminescence of quartz and its application to sandstone petrology. *Contrib. Sedimentol.* **8**, (1978).
- Schieber, J. in *Shales and Mudstones Vol. 1* (eds Schieber, J., Zimmerle, W., & Sethi, P.) 187–215 (Schweizerbart, Stuttgart, 1998).
- Ettensohn, F. R. et al. in *Devonian of the World Vol. 2* (eds McMillan, N. J., Embry, A. F. & Glass, D. J.) 323–345 (Canadian Society of Petroleum Geologists, Calgary, 1988).
- Riciputi, L. R., Paterson, B. A. & Ripperdan, R. L. Matrix effects in the analysis of light (S, C, O, H) isotope ratios by SIMS. *Int. J. Mass Spectrom. Ion Process.* **178**, 81–112 (1998).
- Blatt, H. Oxygen isotopes and the origin of quartz. *J. Sedim. Petrol.* **57**, 373–377 (1987).
- Blatt, H. & Totten, M. W. Detrital quartz as an indicator of distance from shore in marine mudrocks. *J. Sedim. Petrol.* **51**, 1259–1266 (1981).
- Schieber, J. Distribution and deposition of mudstone facies in the Upper Devonian Sonyea Group of New York. *J. Sedim. Res.* **69**, 909–925 (1999).
- Carroll, A. R., Stephens, N. P., Hendrix, M. S. & Glenn, C. R. Eolian-derived siltstone in the Upper Permian Phosphoria Formation: Implications for marine upwelling. *Geology* **26**, 1023–1026 (1998).
- Parrish, J. T. *Interpreting Pre-Quaternary Climate from the Geologic Record* (Columbia University Press, New York, 1998).
- Wollast, R. & Mackenzie, F. T. in *Silicon Geochemistry and Biogeochemistry* (ed. Aston, S. R.) 39–76 (Academic, London, 1983).
- Füchtbauer, H. *Sedimente und Sedimentgesteine* (Schweizerbart, Stuttgart, 1988).

22. Tyson, R. V. & Pearson, T. H. in *Modern and Ancient Continental Shelf Anoxia* (eds Tyson, R. V. & Pearson, T. H.) 1–24 (Special Publication 58, Geological Society of London, 1991).
23. Calvert, S. E. & Pedersen, T. F. in *Organic Matter: Productivity, Accumulation and Preservation in Recent and Ancient Sediments* (eds Whelan, J. K. & Farrington, J. W.) 231–263 (Columbia University Press, New York, 1992).
24. Keperle, R. C. in *Petroleum Geology of the Devonian and Mississippian Black Shale of Eastern North America* (eds Roen, J. B. & Keperle, R. C.), *USGS Bull.* **1909**, F1–F23 (1993).
25. Round, F. E., Crawford, R. M. & Mann, D. G. *The Diatoms: Biology and Morphology of the Genera* (Cambridge Univ. Press, Cambridge, 1990).
26. Maliva, R. G., Knoll, A. H. & Siever, R. Secular change in chert distribution; a reflection of evolving biological participation in the silica cycle. *Palaios* **4**, 519–532 (1989).
27. Ettensohn, F. R. in *Shales and Mudstones Vol. 1* (eds Schieber, J., Zimmerle, W. & Sethi, P.) 109–128 (Schweizerbart, Stuttgart, 1998).
28. O'Brien, N. R. & Slatt, R. M. *Argillaceous Rock Atlas* (Springer, New York, 1990).
29. Hathon, C., Sibley, D. & Cambray, F. W. *The Origin of the Quartz in Antrim Shale* (Report FE-2346-61, US Department of Energy, 1980).
30. Hallam, A. *Phanerozoic Sea-Level Changes* (Columbia Univ. Press, New York, 1992).

Acknowledgements

This work was supported in part by the Geoscience Research Program, Office of Basic Energy Sciences, US Department of Energy with Oak Ridge National Laboratory, managed by Lockheed Martin Energy Research Corp., the National Science Foundation, and the Donors of the Petroleum Research Fund, administered by the American Chemical Society.

Correspondence and requests for materials should be addressed to J.S. (e-mail: schieber@uta.edu).

Density cycles and an offspring quantity and quality game driven by natural selection

Barry Sinervo*, Erik Svensson*† & Tosha Comendant*

* Department of Ecology and Evolutionary Biology, Earth and Marine Sciences Building, University of California, Santa Cruz, California 95064, USA

A long-standing hypothesis^{1,2} posits that natural selection can favour two female strategies when density cycles. At low density, females producing many smaller progeny are favoured when the intrinsic rate of increase, r , governs population growth. At peak density, females producing fewer, high-quality, progeny are favoured when the carrying capacity, K , is exceeded and the population crashes. Here we report on the first example of a genetic r versus K selection^{3–5} game that promotes stable population cycles in lizards. Decade-long fitness studies and game theory demonstrated that two throat-colour morphs were refined by selection in which the strength of natural selection varied with density. Orange-throated females, r strategists, produced many eggs and were favoured at low density. Conversely, yellow-throated females, K strategists, produced large eggs and were favoured at high density. Progeny size should also be under negative frequency-dependent selection in that large progeny will have a survival advantage when rare, but the advantage disappears when they become common. We confirmed this prediction by seeding field plots with rare and common giant hatchlings. Thus, intrinsic causes of frequency- and density-dependent selection promotes an evolutionary game with two-generation oscillations.

The side-blotched lizard (*Uta stansburiana*) has an annual life history in that it matures in one year but rarely survives more than one reproductive season. Fecundity is high and females lay up to five clutches (2–10 eggs per clutch, mean is 6 eggs) at monthly intervals from March to August at Los Baños Grandes, Merced County, California. Progeny defend territories at hatching. Density-dependent natural selection should be intense: adult density can exceed

850 per hectare and 2,600 hatchlings can be produced per hectare. From 1989 to 1999, we measured lizard density and collected eggs from females. Fitness was assessed by counting the number of progeny that reached maturity⁶. In 1992, we discovered two morphs of females—yellow-throated and orange-throated⁷. From 1993 to 1999, we studied frequency cycles, natural selection and the genetics of throat colour⁶.

On the basis of 5,618 individually marked progeny released during a decade of study, annual progeny survival oscillated with a two-year period, as did the mean egg mass and mean clutch size produced at maturity (Fig. 1a–c). Population density of adult females oscillated with a two-year period (Fig. 1d). Density also oscillated synchronously in an adjacent population (Fig. 1d). We designated years when progeny cohorts had low survival as 'crash' years, because the density of progeny as adults was reduced. Progeny cohorts that matured in crash years laid small clutches of small eggs (Fig. 1c). We designated years when progeny cohorts had high survival as 'boom' years, because the density of progeny as adults increased. Progeny cohorts that matured in boom years laid large clutches of large eggs (Fig. 1c). Population cycles cause density environments of parents and progeny to be diametrically opposite, which should lead to oscillatory natural selection. We observed the predicted two-year oscillation in natural selection acting on the egg mass of clutches 2–5 (Fig. 1e), which is when hatchling density is highest and selection should be most intense⁸. The selection gradients^{8,9}, which measure the strength of natural selection, are among the highest reported in nature^{10–12}. In contrast, selection was weaker on clutch 1 and did not oscillate (Fig. 1e).

Population parameters also exhibited long-term trends unrelated to two-year cycles. For example, adult density rose over the period 1989–1992, but fell thereafter, reflecting a significant long-term quadratic change over a decade ($F_{1,8} = 5.21$, $P = 0.05$, Fig. 1).

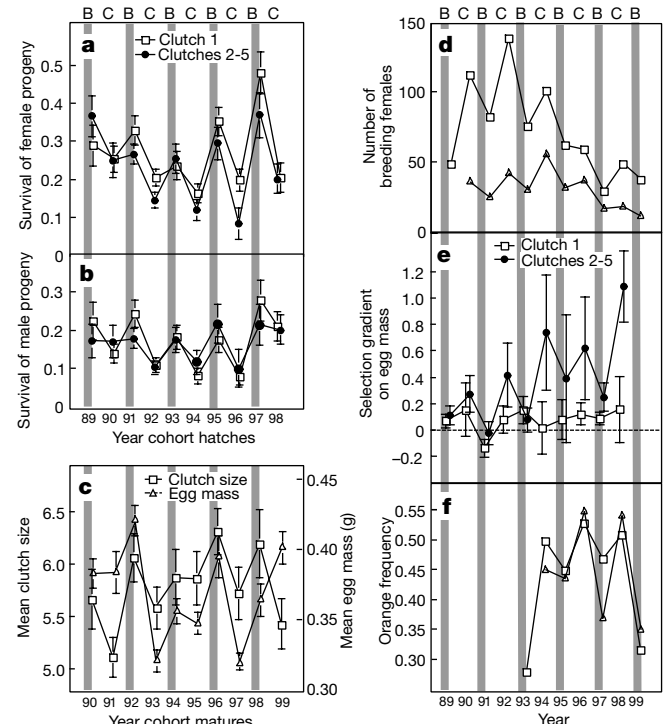


Figure 1 Cycles of demographic parameters and natural selection on egg mass. **a–c**, We observed oscillations of survival of female progeny (**a**), survival of male progeny (**b**), egg mass (triangles), and clutch size (squares) produced by progeny at maturity (**c**). B, boom years (grey bars); C, crash years. **d**, Adult density oscillated in a population where we measured survival and fitness (squares) and in a control population where we only counted density (triangles). **e**, Selection gradients acting on egg size of clutches 2–5 oscillated, but selection gradients on clutch 1 did not oscillate. **f**, Orange frequency oscillated synchronously in two populations. Means $\pm$ s.e.m. are shown.

† Present address: Department of Animal Ecology, Lund University, S-223 62 Lund, Sweden.

Long-term trends were caused by a five fold increase of snake predators¹³ in 1993, which reduced lizard density thereafter. We note that the two-year cycles persisted after an increase in predators, which suggests an intrinsic cause. Before testing for the statistical significance of the two-year cycles¹⁴, we computed residuals to remove the effects of significant long-term trends that were unrelated to two-year cycles (for example, the quadratic trend for survival: $F_{1,34} = 7.44$, $P = 0.01$, the linear trend for selection on clutches 2–5: $F_{1,8} = 6.00$, $P = 0.04$). Temporal autocorrelations were significant for density on both study plots ($r = -0.45$, $r = -0.73$, consensus combined P -value¹⁵ = 0.005), mean density ($r = -0.65$, $P = 0.03$), selection gradients (clutches 2–5: $r = -0.85$, $P = 0.001$) and progeny survival (clutch 1 females: $r = -0.68$, $P = 0.03$; clutch 1 males: $r = -0.87$, $P = 0.001$; clutch 2–5 females: $r = -0.95$, $P = 0.00003$; clutch 2–5 males: $r = -0.79$, $P = 0.006$). Autocorrelations for clutch and egg size were not significant ($r = -0.32$ and $r = -0.31$, $P > 0.38$), but the autocorrelation for total clutch mass was significant (egg mass $\times$ clutch size: $r = -0.65$, $P = 0.03$). The frequency of throat-colour morphs also oscillated on both plots (Fig. 1f; long-term quadratic trend: $F_{1,10} = 10.29$, $P = 0.008$; autocorrelations: $r = -0.84$, $r = -0.72$, $P < 0.01$; ref. 15).

Natural selection on female morphs M was significantly different for clutch sizes s produced on clutch 1 ($M \times s$: $F_{1,257} = 5.531$, $P = 0.02$) and for optimal egg mass m produced on clutches 2–5 ($M \times m$: $F_{1,257} = 6.98$, $P < 0.001$, (m_e)²: $F_{1,257} = 5.18$, $P = 0.02$, Fig. 2a). Thus, orange females laying large clutches had more

progeny that survived to maturity than those laying small clutches. In contrast, yellow females laying large eggs had more progeny that survived to maturity than those laying small eggs. Throat colour was heritable between dams and daughters ($h^2 = 0.48$, $F_{1,116} = 6.92$, $P = 0.001$). Dam's throat colour was also genetically correlated with daughter's clutch size ($G_{t,s} = 1.09$, $P = 0.02$, $N = 168$) and egg mass ($G_{t,m} = -0.84$, $P = 0.05$). This is consistent with the action of correlational selection¹⁶ (Fig. 2a). That is, orange dams produced surviving orange daughters that laid many small eggs. Yellow dams produced surviving yellow daughters that laid fewer large eggs. Morphs differed in clutch size ($s_o = 5.92$ eggs, $s_y = 5.60$ eggs, $F_{1,115} = 4.99$, $P = 0.03$) and egg mass ($m_o = 0.371$ g, $m_y = 0.394$ g, $F_{1,115} = 7.07$, $P = 0.009$) when isolated from other females (for example, no neighbours), reinforcing the view that morphs had distinct life history strategies.

The fitness of each morph in competition with alternative morphs tests for an evolutionarily stable strategy (ESS)¹⁷. The ESS receives the highest fitness pay-off when rare and common. An ESS can thereby invade and eliminate other morphs. The pay-off matrix can be computed from the significant relationship between morph fitness and density of orange and yellow neighbours^{7,18} (Fig. 2b; analysis of covariance, Table 1). When orange females had more orange neighbours their fitness was reduced, but fitness increased with more yellow neighbours. Yellow female fitness was not affected by density of either morph. The pay-off matrix (Table 1) indicated that neither morph was an evolutionary stable state. Both morphs can invade and they may co-exist in an evolutionary stable state.

Even when co-existence is possible, stability must be assessed with population genetic and population dynamic models. We assumed random mating, and that throat colour is due to one locus in which the orange allele is dominant to the yellow allele in females, which is consistent with dam-progeny (above) and sire-progeny pedigrees (B.S. *et al.*, unpublished data). Transmission of morphs across generations is further determined by the frequency dependence of female fitness^{7,18}. Fitness of the i th strategy, $W_i(t)$, is given by the fitness when solitary, $W_{i,sol}$, adjusted by gain or loss in fitness of the i th strategy, G_{ij} , when playing against the j th neighbouring strategy that is found at density, $N_{ij}(t)$:

$$W_i(t) = W_{i,sol} + \sum_j G_{ij} N_{ij}(t) \quad (1)$$

Equation (1) is parameterized by the same relationships used to construct the pay-off matrix (Table 1). We further assumed that progeny survival is density dependent, which was confirmed by a decade of field experiments (regression of survival on density: $F_{1,69} = 7.80$, $P = 0.007$, Fig. 3a). The carrying capacity of orange-throated females ($K_o = 0.70$), which vigorously defended territories, was lower than that of yellow-throated females ($K_y = 1.18$), which were more density tolerant. Recruitment rate of the i th morph at time t was modelled with two coupled logistic equations:

$$\frac{dN_i(t)}{dt} = N_i(t) \times F_i \times \frac{W_i(t)}{\bar{W}} \times \frac{(K_i - N_i(t))}{K_i} \quad (2)$$

Table 1 Fitness pay-offs for rare throat-colour morphs when competing against each common morph

	Population state	
	Orange common	Yellow common
Fitness of rare orange strategy	1	1.61
Fitness of rare yellow strategy	1.60	1

We calculated fitness of rare and common morphs by inserting mean density of each morph, which were observed in boom and crash years, into equation (1)¹⁸. $W_o = 0.98 - 0.84 \times N_{oo} + 0.27 \times N_{oy}$ and $W_y = 1.04 + 0.02 \times N_{yo} - 0.35 \times N_{yy}$ on clutch 1; $W_o = 0.18 - 0.38 \times N_{oo} - 0.14 \times N_{oy}$ and $W_y = 0.22 - 0.03 \times N_{yo} - 0.35 \times N_{yy}$ on clutches 2–5. (Repeated measures ANCOVA of fitness on clutch 1 (C_1) versus clutches 2–5 (C_{2-5}) indicated that fitness response of morphs to density was significantly different: $C \times M \times N_o$: $F_{1,115} = 3.96$, $P = 0.05$, $C_{2-5} \times M \times N_y$: $F_{1,115} = 3.97$, $P = 0.04$.)

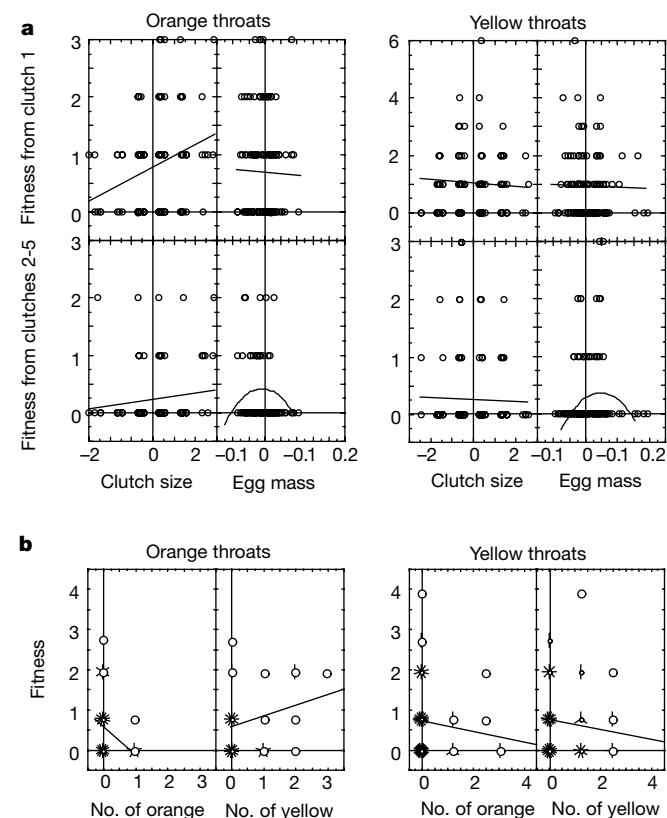


Figure 2 Natural selection on throat-colour morphs. **a**, Natural selection on life history variation differed between morphs. For clutch 1, orange-throated females were favoured to lay larger clutches than yellow-throated females. For clutches 2–5, yellow-throated females were favoured to lay eggs that were 0.08 g heavier on average than orange-throated females. **b**, Orange female fitness was decreased by orange neighbour density but enhanced by yellow neighbour density. Yellow female fitness was not affected by density. We only show pooled data for clutches 1–5, but see Table 1 for repeated measures ANCOVA of fitness on clutch 1 versus clutches 2–5, and equations for fitness. Sample size is indicated by multiple rays on points.

Morphs are coupled through genetics and the frequency dependence of fitness. The intrinsic rate of increase, r , in the standard logistic equation is replaced by F_i , fecundity of i th strategy adjusted by: (1) frequency-dependent fitness (such as equation (1) standardized by mean fitness, $W_i(t)/\bar{W}$), and (2) density-dependent progeny survival of the i th strategy (such as $K_i - N_i(t)/K_i$). Numerical simulations of a discrete logistic ESS model (equation (2)), appropriate for annual side-blotched lizards, confirmed that morph frequency oscillates with a two-year period (Fig. 4). Juvenile survival (for example, $N_i(t+1)/F_i(t)$) also oscillates in the logistic ESS model. Frequency-dependent selection maintains both morphs in an evolutionary stable state, whereas time lags in density-dependent regulation promote stable oscillations. Time lags occur because the evolutionary change in morph frequency (Fig. 1f) is necessarily one generation removed from the action of selection. Thus, the evolutionary lags in our model are analogous to time lags in ecological models that also promote stable cycles¹⁹.

Every other generation, the cycle alternately favours large progeny (crash year) or large clutches of small progeny (boom years). Thus, our model requires that progeny size is under negative frequency-dependent selection. That is, large hatchlings should be selectively advantageous when rare, but disadvantageous when common. We tested this prediction by manipulating the size–frequency distribution of progeny on statistically independent plots. We used experimental variation in egg size, rather than natural variation. A selection study of natural variation in egg size, which differ between morphs, could be confounded with selection on throat-colour alleles or some other correlated trait. However, experimental variation in egg size tests for causes arising from egg size itself²⁰. The progeny size distribution was adjusted by seeding plots with either

20% giant progeny (common giants, $n = 10$ plots) or 10% giant progeny (rare giants, $n = 17$ plots). The remaining frequencies were split between normal-sized and miniaturized progeny. Manipulated progeny are not abnormally large or small²¹. The size range and density of experimental progeny was carefully calibrated to explore the range of variation found in nature^{8,21}. As predicted, strong selection favouring large progeny was observed on plots where giants were rare, and no selection was observed on plots where giants were common (treatment: $F_{1,23} = 4.19$, $P = 0.05$; clutch effect: $F_{1,23} = 2.64$, $P = 0.12$; year: $F_{1,23} = 2.46$, $P = 0.13$, Fig. 3b).

When selection pressure on progeny size is relaxed as on common giant plots, laying more eggs maximizes female fitness. However, females laying more eggs must lay smaller eggs because of the clutch size and egg mass trade-off (that is, genetic correlation, $G_{sm} = -0.91$, $P < 0.01$)²². Orange females lay many small eggs and are only favoured in low-density boom years. In these years, their small offspring are not a fitness liability. Conversely, yellow females are favoured in crash years. In these years, their rare large progeny survive better than small orange progeny.

The alternative throat-colour morph in side-blotched lizards may have arisen as a mutation in an endocrine gene, given the effect of hormones on the expression of orange versus yellow throat colour of lizards²³. This locus will have many pleiotropic effects on physiology^{6,13,22} and behaviour⁷. However, it seems unlikely that all variation in clutch size and egg mass of morphs is due to the pleiotropic effect of a single throat-colour locus. Many loci undoubtedly affect clutch size and egg mass⁵. We have shown (Fig. 1e and Fig. 2a) that correlational selection¹⁶ associated with density cycles is strong enough to form genetic correlations among loci for throat colour, clutch size and egg mass (such as linkage disequilibrium). Selection is disruptive in that it shapes genetic covariation to either side of the clutch and egg size trade-off²². As each strategy becomes better adapted to a different phase of the cycle, oscillations are reinforced. Adaptation of morphs as a result of correlational selection will equilibrate when segregation and recombination erode favourable linkage disequilibrium as fast as it is formed by selection. This results in a chronic selective load on morphs²⁴ that can only be reduced if life-history loci become physically linked to throat-colour loci in a tight linkage group.

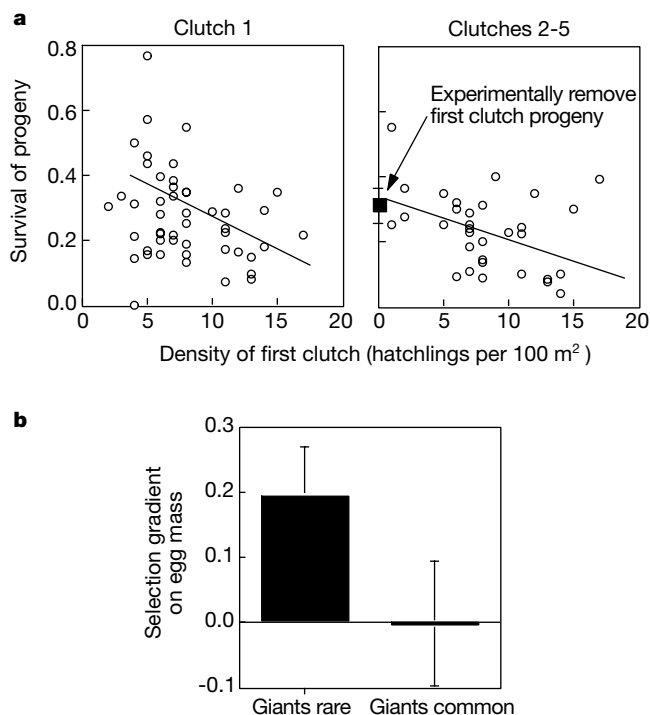


Figure 3 Density-dependent survival and frequency-dependent selection on egg mass. **a**, Survival of progeny from clutches 1 and 2–5 was lower at high density on statistically independent plots. Survival of progeny from clutches 2–5 on experimental plots where clutch 1 progeny had not been previously released (square with s.e.m.) was similar to the regression line intercept for survival of clutches 2–5 progeny on control plots. **b**, Selection acting on progeny size is dependent on frequency distribution of progeny. Positive selection gradients^{8,9} on egg mass were measured on plots with rare giants. No selection (that is, selection gradients were zero) was measured on plots with common giants.

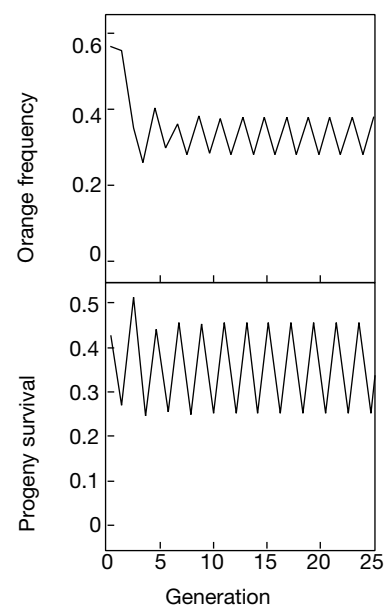


Figure 4 Frequency of orange female morphs and juvenile survival. Data oscillated with a two-year period in simulations of a logistic evolutionarily stable strategy (ESS) model (equation (2)). Top, orange female morphs; bottom, survival of progeny.

The genetic consequences of density-dependent selection cause morphs to refine each other in an endless evolutionary game. The r - and K -strategy model is similar to a hypothesis proposed^{1,2} to explain density cycles in rodents. Our study shows that r -strategists, orange females, are favoured at low density because they produce large clutches of small eggs. The r -strategists cause an overshoot of carrying capacity within a single year, which triggers a population crash. Yellow females, the K -strategists, gain a rare advantage at high density through production of fewer but larger hatchlings. Yellow progeny become larger and better adapted to survive ensuing population crashes, which concomitantly lowers post-crash density of orange. After successive crashes, orange females become better adapted by laying even more small progeny, which concomitantly enhances the overshoot of carrying capacity. Evolutionary and population density time lags ensure that selection will never reach a compromise, but will oscillate between two life-history states. Life-history theory predicts that a single egg size is optimal in a population^{5,21,25,26}. However, it is clear that optimality models are too simplistic when selection arises from density competition. Game theory, which predicts multiple optima for egg size, is more appropriate. □

Methods

Population censuses

In all analyses, female fitness was assessed by tracking survival of individually marked (toe-clipped) progeny to maturity the following year^{6,8,22}. A 500- to 1,500-m diameter area around progeny release sites was searched the following year and emigration does not confound fitness estimates²². Progeny that survived to maturity were unlikely to be missed during daily censuses (March–June)²². We collected sequential clutches from mature female progeny from March to August^{7,13}. Adult density and throat colour⁷ was counted in two populations that were separated by 300 m (ref. 13) of unsuitable adult habitat. Populations exchanged fewer than 5% migrants each year²². Morph frequency varied significantly among years in two populations (Fig. 1f, population where fitness was studied: $\chi^2 = 33.87$, $P < 0.0001$, $n = 102, 113, 110, 108, 49, 61$ and 47; control population: $\chi^2 = 7.13$, $P < 0.01$, $n = 124, 48, 44, 19, 14$ and 68).

Measuring strength of natural selection

Strength of natural selection on egg mass was assessed on statistically independent plots (see below) with selection gradients⁹. Selection gradients are derived from regression of survival on egg mass. Large positive selection gradients reflect strong natural selection that favours large eggs (a steep slope). Standardized selection gradients⁹ were calculated for each plot (1989–1998), while holding hatching-date effects constant^{8,9}.

Significance tests for two-year oscillations

We tested for statistical significance of two-year oscillations after removing the effect of long-term trends (such as decade-long trends) on the yearly means. Residuals with one-year lags (a significant negative correlation for residuals between successive years) indicate two-year oscillations¹⁴ or a significant one-year temporal autocorrelation.

Natural selection on clutch size and egg mass

We measured fitness from the number of progeny that survived to maturity (see above) as a function of female's clutch size and egg mass produced on clutch 1 and clutches 2–5. We tested for significant morph differences in selection using analysis of covariance (ANCOVA)^{7,9} with morph M as a factor, and clutch size (s) or egg mass (m) as covariates. A significant difference in directional selection^{7,9} between morphs was tested with interaction terms^{7,21} ($M \times s$). In addition, we tested for significant stabilizing selection (optimum) by including a term for m^2 . In this model, an effect of $M \times m$ indicates a significant difference in optimal egg mass between morphs^{7,21}.

Local density (N_{ij}) and fitness

The number of neighbouring female morphs was calculated from home range maps^{7,13,17}. Home ranges were computed from the minimum convex polygon circumscribing locations of adult females mapped each spring (1993–1995)^{7,13}. We tested for differences in the fitness response of morphs to density with repeated measures ANCOVA. Fitness of females on clutch 1 versus clutches 2–5 (C) was regressed on a factor for M and covariates for number of orange (N_o) and yellow neighbours (N_y). Differences in morph responses were tested with interaction terms ($M \times C \times N_o$, $M \times C \times N_y$).

Heritability and genetic correlations

In our population censuses (see above), we measured clutch size and egg mass, and scored throat colour of the female parent and her individually marked progeny. We used these field pedigrees to estimate heritability of all traits²². Throat colour was scored as 0 if yellow or 1 orange⁷. Genetic correlations were computed from the reciprocal cross-covariance between parent and progeny traits²² (for example, throat of mother and clutch size or egg

mass of progeny) and from the ordinary covariance between parent–progeny traits.

Experiments on density-dependent survival

During 1989–1998, we released progeny from clutches 2–5 into areas that were either previously seeded with clutch 1 progeny ($n = 35$ control plots) or not seeded with clutch 1 progeny ($n = 18$ treatment plots). Eggs were incubated in the laboratory and progeny were individually marked at hatching; they were then randomly released with respect to dam's territory⁶. Each study plot⁸ consisted of a rock outcropping separated by unsuitable lizard habitat (such as grassland). Dispersal of hatchlings largely occurs within study plots²². Area of plots varied from 311 to 953 m² (mean is 675, s.e.m. = 79). Number of progeny released per study plot varied from 8 to 61 (mean is 28, s.e.m. = 1.3).

Manipulating progeny size–frequency distributions

During 1995–1997, we seeded progeny onto plots where giant progeny were either common (20%, $n = 10$ plots) or rare (10%, $n = 17$ plots) and measured selection gradients on egg size based on survival of progeny to maturity. Remaining frequencies were split between normal-sized and miniaturized progeny. Density was constant within years⁸. Giant eggs were obtained by ablating ovarian follicles²⁷, which yields hatchlings 20% larger than the mean²¹. Miniature eggs were obtained by aspirating 20% of the yolk from freshly laid eggs with sterile syringes^{22,28}, which yields hatchlings 20% smaller than the mean. Normal-sized eggs were poked with sterile syringes but no yolk was removed. Number of progeny released per plot varied from 10 to 41 (mean = 22; s.e.m. = 1.5).

Received 18 May; accepted 30 June 2000.

- Chitty, D. Regulation in numbers through changes in viability. *Can. J. Zool.* **38**, 99–113 (1958).
- Chitty, D. Population forces in the vole and their relevance to general theory. *Cold Spring Harb. Symp. Quant. Biol.* **22**, 277–280 (1960).
- MacArthur, R. H. & Wilson, E. O. *Theory of Island Biogeography* (Princeton Univ. Press, Princeton, 1967).
- Pianka, E. R. On r - and K -selection. *Am. Nat.* **104**, 592–597 (1970).
- Stearns, S. C. *The Evolution of Life Histories* (Oxford Univ. Press, Oxford, 1992).
- Sinervo, B. in *Maternal Effects as Adaptations* (eds Mousseau, T. A. & Fox, C. W.) 288–306 (Oxford Univ. Press, Oxford, 1998).
- Sinervo, B. & Lively, C. M. The rock-paper-scissors game and the evolution of alternative male strategies. *Nature* **380**, 240–243 (1996).
- Svensson, E. & Sinervo, B. Experimental excursions on adaptive landscapes: density-dependent selection on egg size. *Evolution* (in the press).
- Lande, R. & Arnold, S. J. The measurement of selection on correlated characters. *Evolution* **36**, 1210–1226 (1983).
- Boag, P. T. & Grant, P. R. Intense natural selection in a population of Darwin's finches (Geospizinae) in the Galapagos. *Science* **214**, 82–85 (1981).
- Endler, J. A. *Natural Selection in the Wild* (Princeton Univ. Press, Princeton, 1986).
- Gibbs, H. L. & Grant, P. Oscillating selection on Darwin's Finches. *Nature* **327**, 511–513 (1987).
- Sinervo, B. & DeNardo, D. F. Costs of reproduction in the wild: path analysis of natural selection and experimental tests of causation. *Evolution* **50**, 1299–1313 (1996).
- Chatfield, C. *The Analysis of Time Series. An Introduction* (Chapman & Hall, London, 1992).
- Rice, W. R. A consensus combined P-value test and the family-wide significance of component tests. *Biometrics* **46**, 303–308 (1990).
- Brodie, E. E. III Correlational selection for color pattern and antipredator behavior in the garter snake *Thamnophis ordinoides*. *Evolution* **46**, 1284–1298 (1992).
- Maynard Smith, J. *Evolution and the Theory of Games* (Cambridge Univ. Press, Cambridge, 1982).
- Sinervo, B. in *Model Systems in Behavioral Ecology* (ed. Dugatkin, L.) (Princeton Univ. Press, in the press).
- Zeng, Z., Nowierski, R. M., Taper, M. L., Dennis, B. & Kemp, W. P. Complex population dynamics in the real world. Modeling the influence of time varying parameters and time lags. *Ecology* **79**, 2193–2209 (1998).
- Sinervo, B. & Basolo, A. in *Adaptation* (eds Rose, M. R. & Lauder, G. V.) 148–185 (Academic, New York, 1996).
- Sinervo, B., Doughty, P., Huey, R. B. & Zamudio, K. Allometric engineering: a causal analysis of natural selection on offspring size. *Science* **258**, 1927–1930 (1992).
- Sinervo, B. in *Adaptive Genetic Variation in the Wild* (eds Mousseau, T., Sinervo, B. & Endler, J. A.) 41–64 (Oxford Univ. Press, Oxford, 2000).
- Cooper, W. E. & Greenberg, N. in *Hormones, Brain, and Behavior. Biology of the Reptilia* (eds Crews, D. & Gans, C.) 298–422 (Academic, New York, 1992).
- Wallace, C. *Genetic Load, its Biological and Conceptual Aspects* (Prentice Hall, New Jersey, 1970).
- Smith, C. C. & Fretwell, S. D. The optimal balance between size and number of offspring. *Am. Nat.* **108**, 499–506 (1974).
- Brockelman, W. Y. Competition, the fitness of offspring, and optimal clutch size. *Am. Nat.* **109**, 677–699 (1975).
- Sinervo, B. & Licht, P. Proximate constraints on the evolution of egg size, egg number, and total clutch mass in lizards. *Science* **252**, 1300–1302 (1991).
- Sinervo, B. & Huey, R. B. Allometric engineering: an experimental test of the causes of interpopulational differences in locomotor performance. *Science* **248**, 1106–1109 (1990).

Acknowledgements

We thank O. Bjornstad, C. Both, R. Calsbeek, A. Chaine, D. Croll, D. Doak, J. Estes, G. Pogson, N. Janzen, D. Kellogg, P. Lundberg, B. Lyon and J. Vogelzang for comments. Research was supported by NSF grants to B.S. and STINT, Fullbright, and NFR grants to E.S. We thank P. Stadler and R. Schrimpp for permission to work on their land.

Correspondence and requests for materials should be addressed to B.S. (e-mail: sinervo@biology.ucsc.edu).

The role of microbes in accretion, lamination and early lithification of modern marine stromatolites

R. P. Reid*, P. T. Visscher†, A. W. Decho‡, J. F. Stolz§, B. M. Bebout||, C. Dupraz*, I. G. Macintyre¶, H. W. Paerl#, J. L. Pinckney**, L. Prufert-Bebout||, T. F. Stegge# & D. J. DesMarais||

* MGG-RSMAS, University of Miami, 4600 Rickenbacker Causeway, Miami, Florida 33149, USA

† Department of Marine Sciences, University of Connecticut, Groton, Connecticut 06340, USA

‡ School of Public Health, University of South Carolina, Columbia, South Carolina 29208, USA

§ Department of Biological Sciences, Duquesne University, Pittsburgh, Pennsylvania 15282, USA

|| NASA Ames Research Centre, Moffett Field, California 94035, USA

¶ National Museum of Natural History, Smithsonian Institution, Washington DC 20560, USA

Institute of Marine Sciences, University of North Carolina, Morehead City, North Carolina 28557, USA

** Department of Oceanography, Texas A&M University, College Station, Texas 77843, USA

For three billion years, before the Cambrian diversification of life, laminated carbonate build-ups called stromatolites were widespread in shallow marine seas^{1,2}. These ancient structures are generally thought to be microbial in origin and potentially preserve evidence of the Earth's earliest biosphere¹⁻³. Despite their evolutionary significance, little is known about stromatolite formation, especially the relative roles of microbial and environmental factors in stromatolite accretion^{1,3}. Here we show that growth of modern marine stromatolites represents a dynamic balance between sedimentation and intermittent lithification of cyanobacterial mats. Periods of rapid sediment accretion, during which stromatolite surfaces are dominated by pioneer communities of gliding filamentous cyanobacteria, alternate with hiatus intervals. These discontinuities in sedimentation are characterized by development of surface films of exopolymer and subsequent heterotrophic bacterial decomposition, forming thin crusts of microcrystalline carbonate. During prolonged hiatus periods, climax communities develop, which include endolithic coccoid cyanobacteria. These coccoids modify the sediment, forming thicker lithified laminae. Preservation of lithified layers at depth creates millimetre-scale lamination. This simple model of modern marine stromatolite growth may be applicable to ancient stromatolites.

The only known examples of stromatolites presently forming in open marine environments of normal seawater salinity are on the margins of Exuma Sound, Bahamas⁴⁻⁶. Our study focused on well-laminated build-ups at Highborne Cay (76° 49' W, 24° 43' N) as potential analogues of ancient stromatolites extending back to the Precambrian. Highborne Cay stromatolites form in the back reef zone of an algal-ridge fringing reef complex that extends 2.5 km along the eastern shore of the island, facing Exuma Sound⁷. Surface waters have a salinity of 36–37 parts per thousand and are saturated with respect to both aragonite and calcite. Stromatolites form as intertidal and subtidal build-ups, shoreward of the algal ridge. Results reported here pertain to the subtidal stromatolites, which grow in depths of less than 1 m at mean low tide and form ridges and columnar heads up to half a metre high (Fig. 1).

Surfaces of Highborne Cay stromatolites are covered with cyanobacterial mats. Examination of these mats using a variety of integrated geological and microbiological techniques reveals

variations in microbial community structure and composition. Extensive field sampling over a two-year period revealed three mat types, representing a continuum of growth stages with minimal seasonal variability (Fig. 2).

Type 1: About 70% of all mats examined consist of a sparse population of the filamentous cyanobacterium *Schizothrix* sp.⁸. *Schizothrix* filaments are generally vertically orientated and are entwined around carbonate sand grains (Fig. 2a and b).

Type 2: Approximately 15% of mats show development of calcified biofilms, which appear as thin crusts of microcrystalline carbonate (micrite) at the uppermost surface of the mat (Fig. 2c and d). These films are about 20–60 µm thick; they drape over and bridge interstitial spaces between sand grains. Silt-sized carbonate particles, such as tunicate spicules, are commonly embedded in the films. Cyanobacterial filaments are present, but are not abundant in the biofilms, which are comprised mainly of copious amounts of amorphous exopolymer, metabolically diverse heterotrophic microorganisms⁹⁻¹¹ and aragonite needles. Needle-shaped aragonite crystals, approximately 1 µm in length, form spherical aggregates 2–5 µm in diameter and are embedded in the exopolymer matrix (Fig. 2e). Bacteria are abundant and are commonly observed at the edges of the aragonite spherules (Fig. 3). A sparse to moderately dense population of *Schizothrix* underlies the exopolymer biofilm (Fig. 2c).

Type 3: The remaining 15% of mats are characterized by an abundant population of the coccoid cyanobacterium *Solentia* sp. and randomly-orientated *Schizothrix* filaments below a calcified biofilm (Fig. 2f and g). *Solentia* is an endolith, which bores into carbonate sand grains. These bored grains appear grey when viewed in plane polarized light in a petrographic microscope (Fig. 2f), contrasting with the golden-brown colour of unbored grains (Figs 2a and c). The microbored grains are often fused at point contacts and appear 'welded' together (Fig. 2f and g).

The variations in surface mats described above represent changes in microbial community structure and activity in response to intermittent sedimentation. Type 1 mats, characterized by a sparse population of *Schizothrix* filaments, resemble pioneer communities¹², which dominate during periods of sediment accretion. Formation of these mats during intervals of rapid sedimentation is documented by field observations showing that accretion rates of one grain-layer per day produce mats with Type 1 fabrics. The activities of *Schizothrix*, in particular, photosynthetic production of exopolymer, are crucial in the accretion process. Flume studies show that sand grains, which settle from suspension when flow rate is low, adhere to mucous-like exopolymer (B.M.B., unpublished video recordings). These 'trapped' grains are subsequently bound by filaments and exopolymer as *Schizothrix* moves upward to the sediment surface. Populations of diatoms and other eukaryotes are minor to absent in these accreting mats^{8,13,14} indicating that, contrary to previous reports^{15,16}, eukaryotic organisms are not required for the trapping and binding of coarse-grained sediment. Aragonite precipitation is inhibited during this stage through

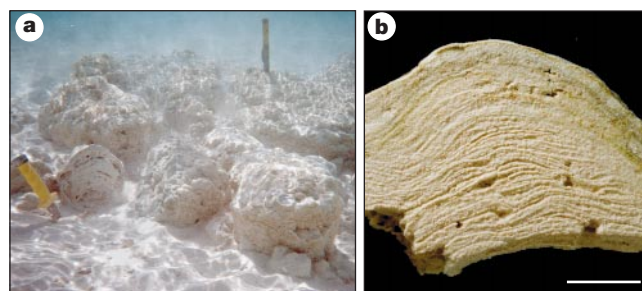


Figure 1 Shallow subtidal stromatolites, Highborne Cay, Bahamas. **a**, Extensive columnar build-ups. **b**, Vertical section showing lamination; scale bar, 2 cm.

calcium ion binding by exopolymer and low-molecular-weight organic acids excreted by *Schizothrix*¹¹.

Type 2 mats represent a more mature^{12,17} surface community characterized by development of a continuous surface film of exopolymer. This mat type develops during quiescent periods when sedimentation ceases and mats begin to lithify. Formation during calm periods is indicated by carbonate silt, such as tunicate spicules, which is commonly entrapped in the surface films but is characteristically lacking in Type 1 mats. Mesocosm manipulations suggest that continuous surface biofilms form in a matter of days. These surface films support heterotrophic activity of both aerobic and anaerobic bacteria^{9,11}, which metabolize the low-molecular-weight organic compounds and the labile fraction of the amorphous exopolymer^{10,11}. Sulphate reduction takes place despite the presence of oxygen at the surface and sulphate-reducing bacteria account for a significant fraction (30–40%) of the organic carbon consumption by the community^{9,10}. This bacterial activity promotes aragonite precipitation as evidenced by microscale observations that high rates of sulphate reduction coincide with micritic crusts¹⁸. In addition, microautoradiography of radiolabelled organic matter shows a close association between bacteria and aragonite needles (H.W.P., unpublished data). The net result of these processes is calcification of the biofilm and formation of a thin micritic crust. When additional carbonate sand is accreted onto the stromatolite, this surface-coating film persists into the subsurface as a nearly continuous thin sheet of micritic cement.

Longer hiatal periods allow formation of Type 3 mats, which are more fully developed than Type 2 mats and include an abundant

population of the coccoid cyanobacterium *Solentia* sp. These *Solentia*-rich mats represent the 'climax' community of the stromatolite system (Fig. 2). Excretion products of *Solentia* and *Schizothrix* support high rates of bacterial respiration^{10,12}. Microscopy and culture experiments have revealed an unusual process of boring and infilling associated with *Solentia*^{19,20}. Boreholes are filled in with aragonite as *Solentia* advances^{19,20}. Moreover, as *Solentia* crosses between grains at point contacts, infilling of the microbored tunnels obliterates grain boundaries and grains become fused together (Fig. 2g). Observations of organic matter in some

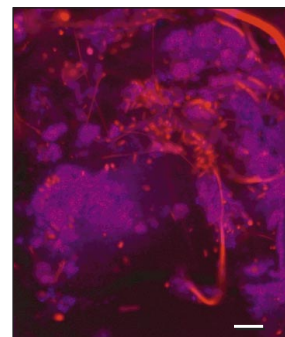


Figure 3 Scanning laser confocal microscope image of a surface biofilm. Bacteria (red fluorescence) are abundant and show an intimate association with carbonate precipitates (blue autofluorescence). Large, uncalcified filament in upper left is *Schizothrix* sp. Sample is stained with propidium iodide; scale bar, 5 μm .

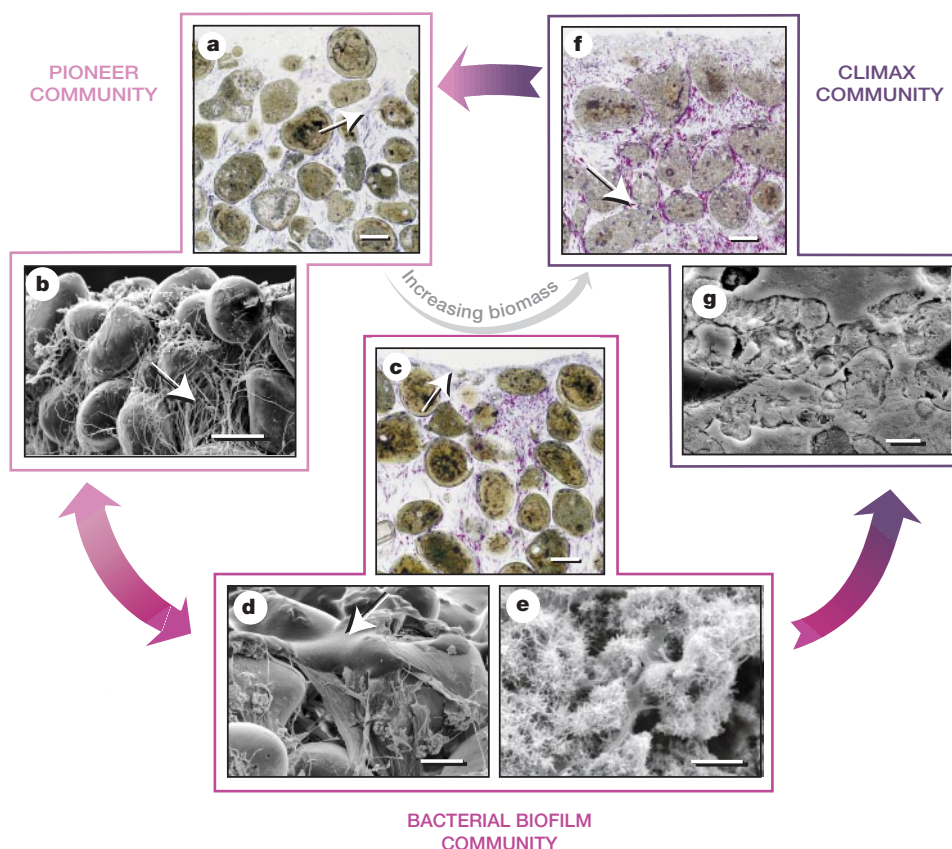


Figure 2 Dominant prokaryotic communities on stromatolite surfaces. Cycling between communities, indicated by large arrows, is a response to intermittent sedimentation (see text). **a, b**, Pioneer community: filamentous cyanobacteria (arrows) bind carbonate sand grains. **c–e**, Bacterial biofilm community: a continuous sheet of amorphous exopolymer (arrows, **c, d**) with abundant heterotrophic bacteria (Fig. 3) forms uppermost surface; aragonite needles precipitate within this surface film (**e**). **f, g**, Climax community: a

surface biofilm overlies filamentous cyanobacteria and endolith-infested grains, which appear grey and are fused (arrow, **f**). Precipitation in tunnels that cross between grains leads to welding (**g**). **a, c, f**, Petrographic thin sections, plane polarized light; cyanobacteria are stained with methylene blue. **b, d, e, g**, Scanning electron microscope images. Scale bars: **a, b, c, f**, 100 μm ; **d**, 50 μm ; **e**, 5 μm ; **g**, 10 μm .

boreholes, together with high sulphate reduction activity in these layers¹⁸ indicates that, as in Type 2 mats, heterotrophic activity may be important in the precipitation process. In contrast to the conventional view that microboring is principally a destructive process^{8,21}, the microboring and infilling processes associated with *Solentia* activity in these mats is a constructive process. This process fuses grains together to create laterally cohesive carbonate crusts. These crusts persist into the subsurface and provide structural support for the growth and long-term preservation of the stromatolite. Field and laboratory studies show that layers of fused microbored grains are formed in periods of weeks to months¹⁹. As *Solentia* is a photosynthetic microorganism, such prolonged periods of microboring activity can only be sustained when this population remains at the surface during long hiatal periods. Even longer hiatal periods result in a community succession to eukaryotic algal communities, which do not form laminated structures^{8,14}.

Laminations in the fossilized part of the stromatolites represent a chronology of former surface mats (Fig. 4). Stromatolite laminae are most easily observed on water-washed, cut surfaces where lithified layers stand out in relief (Fig. 4a). Although lamination is readily apparent in hand samples, it has a subtle expression in petrographic thin sections. Detailed microstructural analyses show, however, that the lithified layers have two distinct petrographic appearances (Fig. 4b). These laminae correspond to (1) thin crusts of micrite, 10–60 μm thick (blue lines in Fig. 4b and c), and (2) layers of fused, microbored grains infested with *Solentia* sp.; these layers are 1–2 mm thick (orange lines in Fig. 4b and d) and underlie micritic crusts. Light microscopy combined with scanning electron microscopy shows that the thin crusts are identical in thickness,

composition and texture to the calcified biofilms described above; they are also similar in thickness to micritic laminae in many ancient stromatolites^{3,22}. In addition, microstructural features of the layers of fused, microbored grains are identical to those formed by the climax community described above.

Lithified layers, which represent former surfaces of mats, show a millimetre-scale distribution. This is indicated by petrographic analyses of the upper several centimetres of 37 stromatolites containing 453 micritic crusts and 174 microbored layers. The micritic crusts form at intervals averaging 1–2 mm. Distances between tops of layers of fused, microbored grains show two modes, one at 2–3 mm and a second at 4–5 mm. Typical marine cements, such as acicular fringes of aragonite, are notably absent during these early stages of growth. Thus, the lithified mats provide initial structural support for the development of laminated build-ups with topographic relief.

To our knowledge, this is the first study to define a specific set of mechanisms that link lamination in marine stromatolites to a dynamic balance between sedimentation, a succession of prokaryotic communities and early lithification. Integration of detailed geological and microbiological analyses of stromatolites in a modern marine system has shown that the structure and composition of surface mats alter in response to intermittent sedimentation and that mats lithify during hiatal periods. Lithification depends on two fundamentally important microbial processes: photosynthetic production by cyanobacteria and heterotrophic respiration by bacteria. A laminated microstructure is formed by precipitation of laterally continuous sheets of micrite in surface biofilms, which are formed during frequent discontinuities in sedimentation. In some

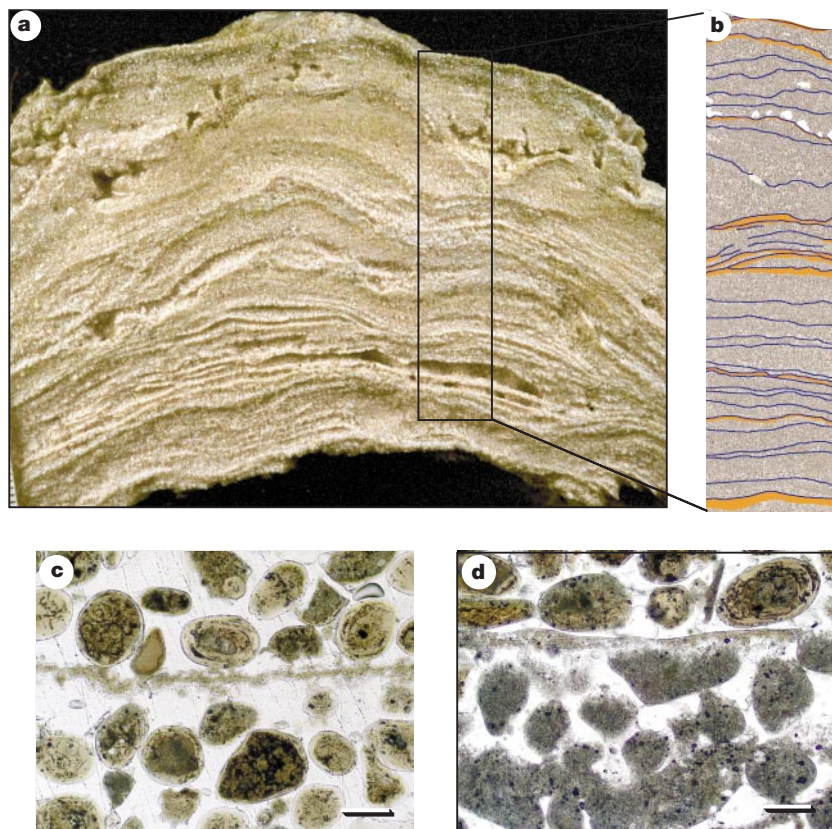


Figure 4 Lamination and microstructure in stromatolite subsurface. Lithified layers representing former surface mats form at 1–2 mm intervals. **a**, Water-washed vertical section showing lithified laminae, which stand out in relief. **b**, Low magnification thin-section photomicrograph of boxed area in **a** showing the distribution of lithified layers. Blue lines represent micritic crusts (**c**); orange lines represent welded, micritized grains

(**d**). **c**, Thin-section photomicrograph of a micritic crust; these crusts represent calcified biofilms (Type 2 mats). **d**, Thin-section photomicrograph of a layer of microbored, fused grains, which underlie a micritic crust; these layers represent former climax communities (Type 3 mats). **c**, **d**, Plane polarized light. Scale bars: **b**, 10 mm; **c**, **d**, 100 μm .

cases, thicker layers of fused grains form below these biofilms in response to microboring activities and precipitation, probably resulting from polymer degradation in boreholes.

These findings provide insight into the role of microbes in stromatolite accretion, lamination and lithification. Although most researchers agree that, "microbial mats and their associated sediments must be lithified early in order to be preserved in the record as stromatolites"²¹, the proposed mechanisms and precise timing of early lithification have been "vigorously debated"²¹. Historically, early lithification was attributed to abiotic processes of submarine cementation^{23,5} or to calcification of cyanobacterial sheaths²⁴ related to photosynthetic activity. More recently, attention has shifted to heterotrophic bacterial decomposition of cyanobacterial sheaths in subsurface, aphotic zones^{25,26}. Although field studies have documented bacterial precipitation of micrite on the sheaths of dead cyanobacteria in the subsurface of laminated microbial mats in tidal flats^{25,27}, these mats do not form fully lithified laminae and stromatolitic build-ups. We argue that growth of laminated microbial structures with topographic relief, such as those that dominated the fossil record for three billion years, depends on penecontemporaneous lithification of surface mats. This lithification process occurs by decomposition of an amorphous matrix of bacterial exopolymer (not sheath material) in the photic zone across the stromatolite surface. Similar processes of precipitation within the amorphous exopolymer matrix of biofilms, rather than on cyanobacterial sheaths, offer an additional mechanism to account for the paucity of preserved microfossils in Precambrian stromatolites, which is typically ascribed to recrystallization and/or rapid degradation of sheaths^{1,26,28}. The potential role that climax microbial communities, functionally equivalent to the endolithic coccoid cyanobacterial communities in modern marine stromatolites, may have played in the growth and lithification of ancient stromatolites remains to be evaluated. □

Methods

This study combined a range of geological, microbial and chemical analyses. An extensive field program was conducted during January and June 1997, and March and August 1998. Physicochemical indices of stromatolite mats were determined *in situ*, primarily with O₂, sulphide, pH needle electrodes (0.8 mm outer diameter)⁹, whereas microstructural, chemical and microbial analyses and incubations were done in the laboratory at the field site and in home institutions. Mat communities and microstructural features were identified using a variety of microscope techniques (light, scanning electron, transmission electron, and scanning laser confocal²⁹) and microbial populations were enumerated using epifluorescence microscopy counts⁹, most-probable number enumerations^{9,10} and molecular phylogenetic techniques. Microbial activities were assessed using depth profiles measured with microelectrodes⁹ and radioisotope incubations using ³H, ¹⁴C and ³⁵S (refs 9, 10 and 30). Heterotrophic activity was also studied with microautoradiography of labelled organic matter uptake³⁰. Microscale distribution of sulphate reduction was assessed using Ag foil coated with ³⁵SO₄²⁻ (ref. 18). Exopolymer distribution and production were evaluated by physical and chemical extractions and ¹⁴C-bicarbonate experiments, respectively¹¹. Other methods used are described elsewhere^{9,10,11,13,29,30}.

Received 17 April; accepted 26 May 2000.

- Grotzinger, J. P. & Knoll, A. H. Stromatolites in Precambrian carbonates: evolutionary mileposts or environmental dipsticks? *Annu. Rev. Earth Planet. Sci.* **27**, 313–358 (1999).
- Walter, M. R. in *Early Life on Earth* (ed. Bengtson, S.) Nobel Symposium Vol. 84, 270–286 (Columbia Univ. Press, New York, 1994).
- Walter, M. R. in *Earth's Earliest Biosphere* (ed. Schopf, J. W.) 187–213 (Princeton Univ. Press, Princeton, 1983).
- Dravis, J. J. Hardened subtidal stromatolites, Bahamas. *Science* **219**, 385–386 (1983).
- Dill, R. F., Shinn, E. A., Jones, A. T., Kelly, K. & Steinen, R. P. Giant subtidal stromatolites forming in normal salinity water. *Nature*, **324**, 55–58 (1986).
- Reid, R. P., Macintyre, I. G., Steneck, R. S., Browne, K. M. & Miller, T. E. Stromatolites in the Exuma Cays, Bahamas: Uncommonly common. *Facies* **33**, 1–18 (1995).
- Reid, R. P., Macintyre, I. G. & Steneck, R. S. A microbialite/algal ridge fringing reef complex, Highborne Cay, Bahamas. *Atoll Res. Bull.* **466**, 1–18 (1999).
- Golubic, S. & Browne, K. M. *Schizothrix bebelinii* sp. nova builds subtidal stromatolites, Lee Stocking Island. *Algolog. Stud.* **83**, 273–290 (1996).
- Visscher, P. T. *et al.* Formation of lithified micritic laminae in modern marine stromatolites (Bahamas): the role of sulfur cycling. *Am. Mineral.* **83**, 1482–1491 (1998).
- Visscher, P. T., Gritzer, R. F. & Leadbetter, E. R. Low-molecular weight sulfonates: a major substrate for sulfate reducers in marine microbial mats. *Appl. Environ. Microbiol.* **65**, 3272–3278 (1999).
- Decho, A. W., Visscher, P. T. & Reid, R. P. Cycling and production of natural microbial exopolymers (EPS) within a marine stromatolite. *Aquat. Microb. Ecol.* (submitted).

- Stal, L. J., van Gernerden, H. & Krumbein, W. E. Structure and development of a benthic microbial mat. *FEMS Microbiol. Ecol.* **31**, 111–125 (1985).
- Pinckney, J. L. & Reid, R. P. Productivity and community composition of stromatolitic microbial mats in the Exuma Cays, Bahamas. *Facies* **36**, 204–207 (1997).
- Seong-Joo, L., Browne, K. M. & Golubic, S. in *Microbial Sediments* (eds Riding, R. E. & Awramik, S. M.) 16–24 (Springer, New York, 2000).
- Awramik, S. M. & Riding, R. Role of algal eukaryotes in subtidal columnar stromatolite formation. *Proc. Natl Acad. Sci. USA* **85**, 1327–1329 (1988).
- Riding, R. in *Biostabilization of Sediments* (eds Krumbein, W. E., Paterson, D. M. & Stal, L. J.) Vol. 84, 183–202 (Oldenburg Univ. Press, Oldenburg, Germany, 1994).
- Van Gernerden, H. Microbial mats: A joint venture. *Mar. Geol.* **113**, 3–25 (1993).
- Visscher, P. T., Reid, R. P. & Bebout, B. M. Microscale observations of sulfate reduction: correlation of microbial activity with lithified micritic laminae in modern marine stromatolites. *Geology* (in the press).
- Macintyre, I. G., Prufert-Bebout, L. & Reid, R. P. The role of endolithic cyanobacteria in the formation of lithified laminae in Bahamian stromatolites. *Sedimentology* (in the press).
- Reid, R. P. & Macintyre, I. G. Microboring versus recrystallization: further insight into the micritization process. *J. Sediment. Res.* **70**, 24–28 (2000).
- Golubic, S., Seong-Joo, L. & Browne, K. M. in *Microbial Sediments* (eds Riding, R. E. & Awramik, S. M.) 57–67 (Springer, New York, 2000).
- Bertrand-Sarfati, J. in *Stromatolites, Developments in Sedimentology* (ed. Walter, M. R.) Vol. 20, 251–259 (Elsevier, New York, 1976).
- Logan, B. W. *Cryptozoon* and associate stromatolites from the Recent, Shark Bay, Western Australia. *J. Geol.* **69**, 517–533 (1961).
- Monty, C. L. V. in *Stromatolites, Developments in Sedimentology* (ed. Walter, M. R.) Vol. 20, 193–250 (Elsevier, New York, 1976).
- Chafetz, H. S. & Buczynski, C. Bacterially induced lithification of microbial mats. *Palaios* **7**, 277–293 (1992).
- Bartley, J. K. Actualistic taphonomy of cyanobacteria: implications for the Precambrian fossil record. *Palaios* **11**, 571–586 (1996).
- Krumbein, W. E., Cohen, J. & Shilo, M. Solar Lake (Sinai). 4. Stromatolitic cyanobacterial mats. *Limnol. Oceanogr.* **22**, 635–656 (1977).
- Walter, M. R., Bauld, J., DesMarais, D. J. & Schopf, J. W. in *The Proterozoic Biosphere* (eds Schopf, J. W. & Klein, C.) 335–338 (Cambridge Univ. Press, Cambridge, 1992).
- Decho, A. W. & Kawaguchi, T. Confocal imaging of *in situ* natural microbial communities and their extracellular polymeric secretions (EPS) using nanoplast resin. *BioTechniques* **27**, 1246–1252 (1999).
- Pael, H. W., Bebout, B. M., Joye, S. B. & DesMarais, D. J. Microscale characterization of dissolved organic matter production and uptake in marine microbial mat communities. *Limnol. Oceanogr.* **38**, 1150–1159 (1993).

Acknowledgements

D. A. Dean prepared petrographic thin sections; T. Kawaguchi provided confocal images; N. Pinel analysed layer distribution. Numerous students assisted in field and laboratory studies. Logistical field support was provided by the crew of the RV *Calanus* and Highborne Cay management. This study is a contribution to the Research Initiative on Bahamian Stromatolites Project and International Geological Correlation Project Biodegradation of Microbial Buildups. Funding for this research was provided by the US National Science Foundation, Ocean Sciences Division and NASA's Exobiology Program. Correspondence and requests for materials should be addressed to R.P.R. (e-mail: preid@rsmas.miami.edu).

Ant-like task allocation and recruitment in cooperative robots

Michael J. B. Krieger*†, Jean-Bernard Billeter‡ & Laurent Keller*

* Institute of Ecology, University of Lausanne, BB, 1015 Lausanne, Switzerland
‡ Laboratoire de micro-informatique, Swiss Federal Institute of Technology, 1015 Lausanne, Switzerland

One of the greatest challenges in robotics is to create machines that are able to interact with unpredictable environments in real time. A possible solution may be to use swarms of robots behaving in a self-organized manner, similar to workers in an ant colony^{1–5}. Efficient mechanisms of division of labour, in particular series-parallel operation and transfer of information among group members⁶, are key components of the tremendous ecological success of ants^{7,8}. Here we show that the general principles regulating division of labour in ant colonies indeed allow the design of

† Present address: Department of Entomology and Department of Microbiology, University of Georgia, Athens, Georgia 30602, USA.

flexible, robust and effective robotic systems. Groups of robots using ant-inspired algorithms of decentralized control techniques foraged more efficiently and maintained higher levels of group energy than single robots. But the benefits of group living decreased in larger groups, most probably because of interference during foraging. Intriguingly, a similar relationship between group size and efficiency has been documented in social insects^{9–11}. Moreover, when food items were clustered, groups where robots could recruit other robots in an ant-like manner were more efficient than groups without information transfer, suggesting that group dynamics of swarms of robots may follow rules similar to those governing social insects.

We chose foraging as the task assigned to robots to investigate whether ant-derived algorithms allow several robots to achieve tasks in a cooperative manner because foraging efficiency is a key factor influencing colony productivity^{7,12}, foraging is the task for which mechanisms of task allocation are best understood^{13,14}, and foraging is the task that has received the greatest attention in cooperative robotics⁵. The essential components of social organization governing ant colonies were used to program swarms of robots foraging from a central 'nest' (Fig. 1; see electronic movie, Supplementary Information).

First, 'colony'-level information about colony energy was accessed and updated by the individual robots. While in the nest, the robots were informed of the colony energy by radio messages from a control station. This simulated colony-level information that social insects obtain by other means under natural conditions¹⁵. No further information on colony energy was provided to them during foraging missions. Upon return to the nest, each robot renewed its energy reserves (which corresponded to a decrease in colony energy) and unloaded the food item collected into a basket (if the foraging trip was successful), thereby increasing colony energy. This mimics a natural situation where ants share through trophallaxis food that they stored in their crops (social organ)¹⁵. Second, the robots were programmed to avoid each other to minimize collisions and avoid negative interactions. This directly follows from the observation that ants actively regulate their rates of encounters¹⁶ and because foraging ants, like robots, cannot occupy the same space simultaneously. Third, individual variation in the tendency to perform a task was implemented in the robots. This directly models the different individual stimulus thresholds for task allocation that are observed in ants and bees^{13,17–19}. Thus, robots did not leave the nest simultaneously, but only when the colony energy dropped below the pre-set threshold of a particular robot. The individual stimulus thresholds varied linearly, ranging between 75 and 100% of

the initial colony energy. The initial colony energy was proportional to the number of robots per colony. This resulted in good modulation of the number of individuals engaged in the two possible activities: staying in the nest (inactive) and foraging (active). Finally, for some trials robots were programmed to recruit another robot when they identified a resource-rich area, thus mimicking recruitment behaviour observed in many ant species¹⁵. Hence, the robots had generalized information about the overall colony energy, ways to avoid interfering with one another in space and time, and, in the last experiment, the ability to transfer useful information to others.

To determine the relationships between group size and efficiency, we compared groups of 1, 3, 6, 9 and 12 robots in 3 sets of experiments: uniform food distribution without recruitment; clustered food distribution without recruitment; and clustered food distribution with recruitment. Relative colony energy (expressed as the mean energy per robot) was recorded every 5 min during a 30-min trial period. For the analyses we averaged the last three recordings, as they represented most clearly the differences among experiments and the effect of robot number. The use of more than the last three measures did not affect the outcome of the statistical analyses.

In experiments with uniform food distribution, there was a significant effect of group size (analysis of variance (ANOVA): $F_{4,35} = 4.69$; $P = 0.004$) on the average relative colony energy (Fig. 2). A posteriori comparisons (Fisher's procedures for learning systems design (PLSD) tests) showed that this difference was mainly due to single robots and groups of 12 robots having lower foraging efficiency than groups of intermediate size (differences were significant at $P < 0.05$ between groups 1 and 3, and between groups 12 and 3, 12 and 6, and 12 and 9). Groups of intermediate size performed better when food items were uniformly distributed, which is probably due to a trade-off between the positive and negative effects of robot-robot interactions. Because robots were programmed to avoid each other, groups of robots probably exhibited a more efficient coverage of the foraging arena than single robots. When the number of robots increased, however, negative interactions among robots (here defined as robots interfering with each other when trying to perform a task) also increased. In particular, larger groups of robots more frequently suffered from negative interference at the entrance of the nest, resulting in robots being slower in leaving and returning to the colony. In addition, food items were more frequently seized by two robots in larger groups. Overall, the proportion of time spent in such negative interactions was 0, 0.6, 0.5, 1.5 and 2.0% for 1, 3, 6, 9 and 12 robots,

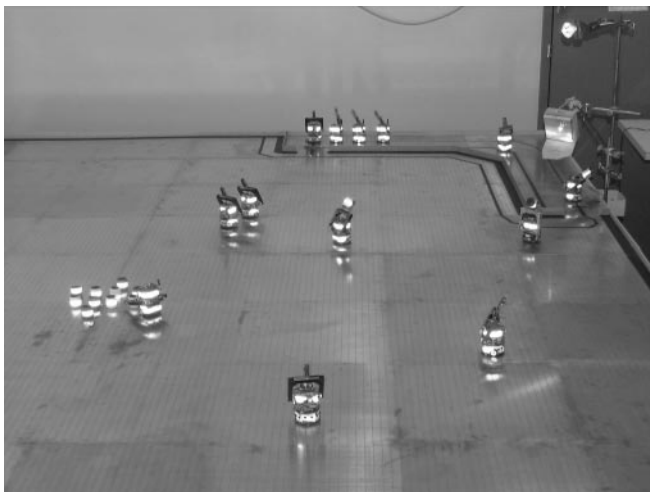


Figure 1 Khepera robots foraging and collecting food items which are then transported back to the nest.

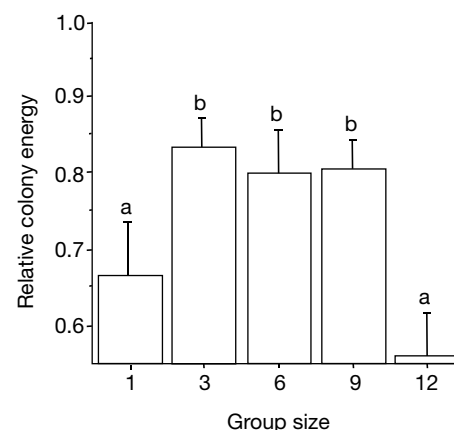


Figure 2 Relative colony energy (per robot) (mean $\pm$ s.e.) in groups of 1, 3, 6, 9 and 12 robots with uniform food distribution and no recruitment. Mean values with different lower-case letters differed significantly ($P < 0.05$, Fisher's PLSD tests).

respectively (ANOVA: $F_{4,35} = 4.43$; $P = 0.005$). A post hoc analysis (Fisher's PLSD, 5%) revealed that negative interactions were more common in groups of 12 robots than in teams of 1, 3 and 6 robots, as well as in groups of 9 compared with 1. Although negative interactions accounted for only a limited amount of the robot's time budget, they significantly affected performance because they occurred predominantly in the vicinity of food items and nest entrance, hence directly affecting the localization and memorization of nest entrance and food items, respectively.

In the second set of experiments with clustered food distribution and no recruitment, we found a similar but nonsignificant relationship (ANOVA: $F_{4,35} = 1.11$; $P > 0.05$) between groups size and colony energy (Fig. 3). The smaller effect of group size in this experiment might have resulted from the decreased benefit provided by uniform coverage of the foraging arena when food items were clustered rather than uniformly distributed. Because robots were programmed to memorize and return to areas where they encountered more than one food item (that is, when they found clustered food items, see Methods), foraging efficiency should have been higher when food was clustered rather than uniformly distributed. Relative colony energy was indeed 10.9% higher when food was clustered (two-way ANOVA, effect of food distribution: $F_{1,70} = 8.69$; $P = 0.004$), a result consistent with the idea that robots were able to relocate patches where they had found more than one food item. The effect of group size was again significant ($F_{4,70} = 4.80$; $P = 0.002$) and the interaction between food item distribution and group size not significant ($F_{4,70} = 0.89$; $P > 0.05$).

The third set of experiments tested for potential benefits conferred by information transfer among the robots. In these experiments food was clustered, as in the second set of experiments, but robots could recruit another robot (see Methods) after their return to the nest. Recruitment significantly increased foraging efficiency compared with experiments where robots could not communicate (Fig. 4), as shown by the overall 9.4% higher mean relative colony energy (two-way ANOVA: effect of information transfer; $F_{1,56} = 9.49$; $P = 0.003$; only groups with two or more robots were considered in the analysis because no information transfer can occur when robots are alone). As in the previous experiments, colony energy was also influenced by group size ($F_{3,56} = 7.70$; $P < 0.001$), but there was no significant interaction between group size and whether or not robots could transfer information ($F_{3,56} = 0.16$; $P = 0.923$).

A similar relationship between group size and group efficiency has been documented in social insects. For example, a detailed study of the wasp *Polybia occidentalis* showed that per capita output

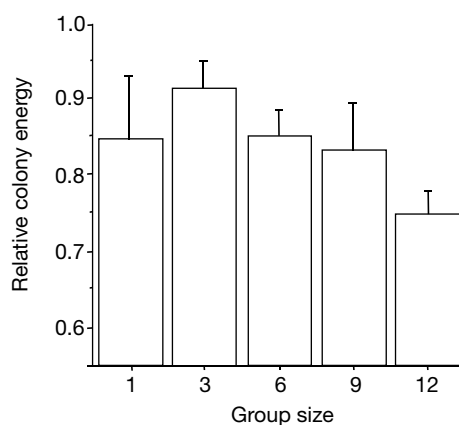


Figure 3 Relative colony energy (per robot) (mean $\pm$ s.e.) in groups of 1, 3, 6, 9 and 12 robots with clustered food distribution and no recruitment.

increases with group size⁹ but levels off in large colonies in a manner similar to that observed with our experimental robots. Decreased individual worker efficiency in large colonies has also been documented in several ant species^{10,11}. One possible reason for this effect is increased rates of negative interference among group members in larger colonies, as has been suggested by theoretical models²⁰ and shown in our experiments with robots. Notably, ants have been shown actively to regulate the rate of encounters as a function of colony size, perhaps to limit the rate of negative interferences at high density and optimize task allocation¹⁶.

Our results show that an ant-inspired system of task allocation based on variation in individual stimulus thresholds provides a simple, robust and efficient way to regulate activity within teams of robots (in our case, collecting food items when robots have no initial information on the environment and group size). Furthermore, relatively complex tasks can be efficiently performed by relatively simple (and autonomous) robots regulated in a decentralized way. This has important implications in robotics, particularly in situations where agents must perform tasks in stochastic environments and where risks of system failures must be avoided, for example during missions on Mars or other planets. A further advantage of controlling robots with individual stimulus thresholds is that it provides great flexibility with regard to team size. As shown in our experiments, single as well as groups of robots can be controlled with the same threshold mechanism. This has also important implications for robotics because it allows for an increase in the number of robots in a system without modifying the control program of individual robots. Likewise, this permits a swarm of robots to continue to work efficiently in case of the number of robots being reduced by individual failures or by allocation of some robots to other work areas or tasks.

Our experiments also show that, similar to a pattern observed in social insects, groups where robots could recruit other robots in an ant-like manner were significantly more efficient at exploiting clustered resources than groups without information transfer. The implementation of an ant-like mechanism of recruitment offers an interesting perspective in robotics when tasks to be performed vary in an unpredictable manner in space and/or time. In conclusion, this study illustrates that similar self-organizing behaviours emerge from groups of agents, be they robots or actual insects, and shows that transfer of knowledge about social insects to the field of intelligent system design is indeed possible. □

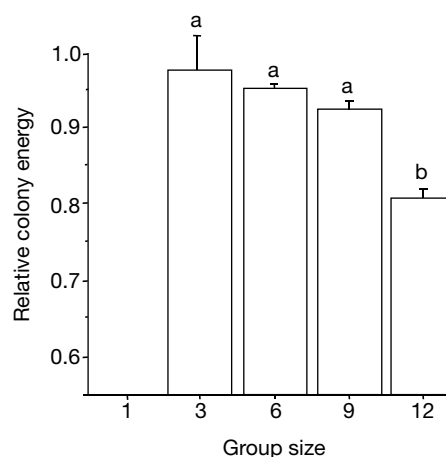


Figure 4 Relative colony energy (per robot) (mean $\pm$ s.e.) in groups of 3, 6, 9 and 12 robots with clustered food distribution and recruitment (no experiments were conducted with single robots as at least 2 robots are necessary for recruitment to occur). Mean values with different lower case letters differed significantly ($P < 0.05$, Fisher's PLSD tests).

Methods

Robots

Detailed information on the robots and programs used are available on request and general information on the system is available on the movie (see Supplementary Information). Essentially, we used Khepera miniature mobile robot designed as a research tool at the 'Laboratoire de micro-informatique' of the Swiss Federal Institute of Technology at Lausanne²¹ (now produced by K-Team SA). Three additional modules, a gripper turret, a custom-made detection module and a radio turret were added (Fig. 1).

Experiments

The experiments were carried out in a 9.24 m² arena surface covered with copper strips alternatively connected to the two poles of a direct current power supply. The food items consisted of small plastic cylinders (3 × 3 cm) with narrow strips of infrared reflecting tape. Depending on the experiment, they were either grouped in two patches (mean distance to nest 1.94 ± 0.06 m, *n* = 26) or placed singly at six locations (mean distance to nest 1.98 ± 0.71 m, *n* = 6). Food items were replaced immediately after being collected by a robot. Robots were programmed so that they randomly foraged in the arena until they either collected a food item or used their energy reserve (autonomy ~4 min), at which time they returned to the nest. Robot energy consumption was low when they were immobile, increased linearly with speed, and was maximal when robots carried a food item.

Once they collected a food item, robots searched the vicinity (within a radius of 7.5 cm, during ~30 s) to determine whether other food items were present. If they found another food item the robots recorded the return path to the nest (through integration of wheel movements) and used this information to return to the patch on the next foraging trip. To allow robots to locate and return to the nest after their foraging trips we placed a lamp close to the nest entrance, allowing robots to use visual cues (as real ants typically do¹⁵) to home-in on the nest.

In the third set of experiments, robots were able to recruit another robot if they had found a second food item (which was generally the case when food was clustered). Once back in the nest, the successful robot recruited the robot at the head of the waiting line in the nest. This experiment investigated potential benefits provided by active recruitment and information transfer^{22,23} and was directly inspired by the process of tandem recruiting behaviour that occurs in ants^{24,25}. The leader robot's speed was lowered during tandem walking to ensure successful arrival of both robots at the food patch which mimics the natural situation in the ants^{24,25}. Once arrived at the food patch, the leader robot decoupled the tandem allowing the follower to begin foraging.

The three sets of experiments were repeated 8 times for each group size (112 experiments in total). To test the effect of group size, we performed one-way ANOVAs²⁶. Comparisons between sets of experiments were performed with two-way ANOVAs to control for the effect of group size.

Received 24 February; accepted 4 July 2000.

- Kube, C. R. & Zhang, H. Collective robotics: From social insects to robots. *Adapt. Behav.* **2**, 189–218 (1994).
- Beckers, R., Holland, O. E. & Deneubourg, J. L. in *Proceedings of the Fourth Workshop on Artificial Life* (eds Brooks, R. & Meas, P.) 181–189 (MIT Press, Boston, 1994).
- Deneubourg, J. L. et al. in *Proceedings of the first International Conference on Simulation of Adaptive Behavior. From Animals to Animals* (eds Meyer, J.-A. & Wilson, S. W.) 181–189 (MIT Press, Boston, 1990).
- Bonabeau, E., Dorigo, M. & Theraulaz, G. *Swarm Intelligence. From Natural to Artificial Systems* (Oxford Univ. Press, Oxford, 1999).
- Cao, Y. U., Fukunaga, A. S. & Kahng, A. B. Cooperative mobile robotics: Antecedents and directions. *Autonom. Robots* **4**, 7–27 (1997).
- Detrain, C., Deneubourg, J. L. & Pasteels, J. M. in *Information Processing in Social Insects* (eds Detrain, C., Deneubourg, J. L. & Pasteels, J. M.) 537–547 (Birkhäuser, Basel, 1999).
- Oster, G. & Wilson, E. O. *Caste and Ecology in the Social Insects* (Princeton Univ. Press, Princeton, 1978).
- Wilson, E. O. *Success and Dominance in Ecosystems: The Case of the Social Insects* (Ecology Institute, Oldendorf/Luhe, Germany, 1990).
- Jeanne, R. L. & Nordheim, E. V. Productivity in a social wasp. Per capita output increases with swarm size. *Behav. Ecol.* **7**, 43–48 (1996).
- Tschinkel, W. R. Sociometry and sociogenesis of colonies of the fire ant *Solenopsis invicta* during one annual cycle. *Ecol. Mon.* **63**, 427–457 (1993).
- Sundström, L. Sex allocation and colony maintenance in monogynous and polygynous colonies of *Formica truncorum* (Hymenoptera: Formicidae): the impact of kinship and mating structure. *Am. Nat.* **146**, 182–201 (1995).
- Anderson, C. & Ratnieks, F. L. W. Task partitioning in insect societies. I. Effect of colony size on queueing delay and colony ergonomic efficiency. *Am. Nat.* **154**, 521–535 (1999).
- Robinson, G. E. Regulation of division of labor in insect societies. *Annu. Rev. Entomol.* **37**, 637–665 (1992).
- Seeley, T. D. Honey bee colonies are group-level adaptive units. *Am. Nat.* **150**, S22–S41 (1997).
- Hölldobler, B. & Wilson, E. O. *The Ants* (Springer, Berlin, 1990).
- Gordon, D. M., Paul, R. E. & Thorpe, K. What is the function of encounter patterns in ant colonies? *Anim. Behav.* **45**, 1083–1100 (1993).
- Robinson, G. E. Modulation of alarm pheromone perception in the honey bee: evidence for division of labour based on hormonally regulated response thresholds. *J. Comp. Physiol.* **160**, 613–619 (1987).
- Robinson, G. E., Fahrback, S. E. & Winston, M. L. Insect societies and the molecular biology of social behavior. *Bioessays* **19**, 1099–1108 (1997).
- Bonabeau, E., Theraulaz, G. & Deneubourg, J. L. Quantitative study of the fixed threshold model for

- the regulation of division of labour in insect societies. *Proc. R. Soc. Lond. B* **265**, 327–332 (1996).
- Pacala, S. W., Gordon, D. M. & Godfray, H. C. J. Effects of social group size on information transfer and task allocation. *Evol. Ecol.* **10**, 127–165 (1996).
- Mondada, F., Franzi, E. & Jenne, P. in *Proceedings of the Third International Symposium on Experimental Robotics 1993* (eds Yoshikawa, T. & Miyazaki, E.) 501–513 (Springer, Berlin, 1994).
- Wilson, E. O. *The Insect Societies* (Harvard Univ. Press, Cambridge, Massachusetts, 1971).
- Seeley, T. D. *The Wisdom of the Hive. The Social Physiology of Honey Bee Colonies* (Harvard Univ. Press, Cambridge, Massachusetts, 1995).
- Hölldobler, B., Möglich, M. & Maschwitz, U. Communication by tandem running in the ant *Camponotus sericeus*. *J. Comp. Physiol.* **90**, 105–127 (1974).
- Möglich, M., Maschwitz, U. & Hölldobler, B. Tandem calling: a new kind of signal in ant communication. *Science* **186**, 1046–1047 (1974).
- Sokal, R. R. & Rohlf, F. J. *Biometry. The Principles and Practice of Statistics in Biological Research* 3rd edn (Freeman, San Francisco, 1995).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>).

Acknowledgements

We thank J. D. Nicoud, E. Franzi, A. Guignard, A. Martinol, P. Maechler and K-Team SA for their sustained help, as well as A. Buchan, M. Chapuisat, T. Clutton-Brock, J.-L. Deneubourg, D. Gordon, B. Jeanne, J. Parker, F. Ratnieks, D. Queller, M. Reuter and K. Ross for comments on the manuscript. This work was funded by the 'Fonds UNIL-EPFL 1997', the 'Stiftung für wissenschaftliche Forschung' and several grants from the Swiss NSF.

Correspondence and requests for materials should be addressed to L.K. (e-mail: Laurent.Keller@ie-zea.unil.ch).

The relative metabolic demand of inhibition and excitation

Daniel Waldvogel*, Peter van Gelderen†, Wolf Muellbacher*‡, Ulf Ziemann*‡, Ilka Immisch* & Mark Hallett*

* Human Motor Control Section, † In Vivo NMR Research Center, NINDS, NIH, Building 10, 10 Center Drive, Bethesda, Maryland 20892, USA

By using the (¹⁴C)2-deoxyglucose method¹, inhibition has been shown to be a metabolically active process at the level of the synapse^{2,3}. This is supported by recent results from magnetic resonance spectroscopy that related the changes in neuroenergetics occurring with functional activation to neurotransmitter cycling⁴. However, inhibitory synapses are less numerous and strategically better located than excitatory synapses, indicating that inhibition may be more efficient, and therefore less energy-consuming, than excitation. Here we test this hypothesis using event-related functional magnetic resonance imaging in volunteers whose motor cortex was inhibited during the no-go condition of a go/no-go task, as demonstrated by transcranial magnetic stimulation. Unlike excitation, inhibition evoked no measurable change in the blood-oxygenation-level-dependent signal in the motor cortex, indicating that inhibition is less metabolically demanding. Therefore, the 'activation' seen in functional imaging studies probably results from excitation rather than inhibition.

Anatomical studies have shown that the cortex consists of about 70–85% spiny neurons (mostly pyramidal cells) and about 15–30% non-spiny non-pyramidal cells^{5,6}. Spiny neurons are excitatory, whereas non-spiny cells constitute different classes of inhibitory interneurons. The relative number of synapses is also different—that is, on average there is about one inhibitory synapse for every five to six excitatory synapses⁷. Inhibitory synapses are located on or near the soma of pyramidal cells as well as on dendritic shafts, whereas excitatory synapses are mainly located on dendritic spines⁸.

‡ Present addresses: Neurological Hospital of Vienna, Riedelgasse 5, 1130 Vienna, Austria (W.M.); Department of Neurology, J.W. Goethe University, Theodor-Stern Kai 7, 60580 Frankfurt am Main, Germany (U.Z.).

Because of their strategically superior location, inhibitory synapses might be more efficient than excitatory synapses, as has recently been demonstrated in the striatum⁹. The difference in number and efficiency between excitatory and inhibitory synapses could lead to lower metabolic demand during inhibition compared with excitation.

We investigated this hypothesis using the recently developed technique of event-related functional magnetic resonance imaging¹⁰ (fMRI), using an experimental protocol that had been shown to either excite or inhibit the primary motor cortex reproducibly. Using transcranial magnetic stimulation (TMS), Leocani and co-workers¹¹ demonstrated that the primary motor cortex was inhibited under the no-go condition of a go/no-go protocol. To ensure that the volunteers who were later studied by fMRI did indeed inhibit their motor cortex under the no-go condition, we performed two TMS experiments. During the first experiment, we chose a single-pulse technique that allowed the same set-up that was later used in the fMRI. In a two-choice reaction time task, go and no-go visual stimuli were presented randomly. Subjects were instructed to press a button as quickly as possible only when the go signal appeared on the screen. At intervals between 200 and 500 ms after either instruction, a TMS pulse was applied to the motor cortex contralateral to the response hand. The peak-to-peak amplitude of the motor evoked potentials (MEPs) in the prime mover muscle was compared with the average of the baseline MEP before and after the reaction time experiment. Six of the eight volunteers screened showed inhibition in the no-go condition. On average, inhibition was strongest at 200 and 300 ms after the instruction, but even at the last time point (500 ms), MEPs were still only 80% of baseline (Fig. 1).

It is conceivable that the decreased MEP size was not due to inhibition but rather to 'disfacilitation', as it is known that task-related changes in attention might influence response size¹². Because we compared the MEP amplitude after the no-go command with the MEP amplitude measured at baseline before and after the go/no-go instructions, there was no reason to assume a facilitating effect on the baseline MEPs. However, to strengthen our argument further, as well as to prove that inhibition was not at the brainstem or spinal level, we performed a second TMS experiment using the double-pulse technique^{13,14}. The experiment, which was technically satisfactory in two of four subjects, demonstrated increased intracortical inhibition following the no-go stimulus.

The single-event fMRI study used the same protocol as the first TMS experiment. We used the PRESTO sequence¹⁵ on a 1.5-T GE scanner (GE Medical Systems) which allowed us to acquire 12 slices with an isotropic voxel resolution of 3.75 mm in 1 s. The field of view was chosen to include the primary motor area and the pre-supplementary motor area (preSMA). This latter area has been shown to be involved in decision making, programming and updating motor programs^{16–18}, and therefore was expected to show activation with both the go and no-go conditions, as was

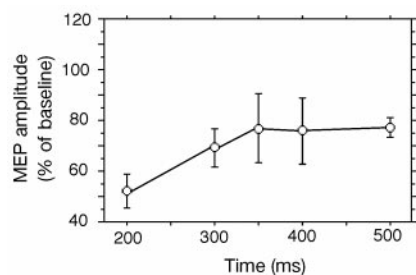


Figure 1 Motor evoked potentials in the no-go condition. The x axis denotes the time after the no-signal; the y axis the averaged motor evoked potential (MEP) size of all six volunteers (± 1 s.e.), expressed as a percentage of the individual baseline MEP size of the flexor digitorum communis muscle.

previously described with a delayed reaction task¹⁹. Figure 2 shows the averaged haemodynamic response in both the primary motor cortex and preSMA for all pixels that passed the Bonferroni corrected threshold of $P < 0.05$ in either condition. While there is similar activation in the preSMA with both the go and no-go conditions, there is no significant change from baseline for the no-go task in the primary motor area.

The haemodynamic response curves shown represent the blood-oxygen-level-dependent (BOLD) MRI signal over time, which correlates with cerebral blood flow and therefore metabolism^{20–24}. This correlation is not linear, and an exact quantification of the metabolic demand on the basis of the BOLD signal is therefore

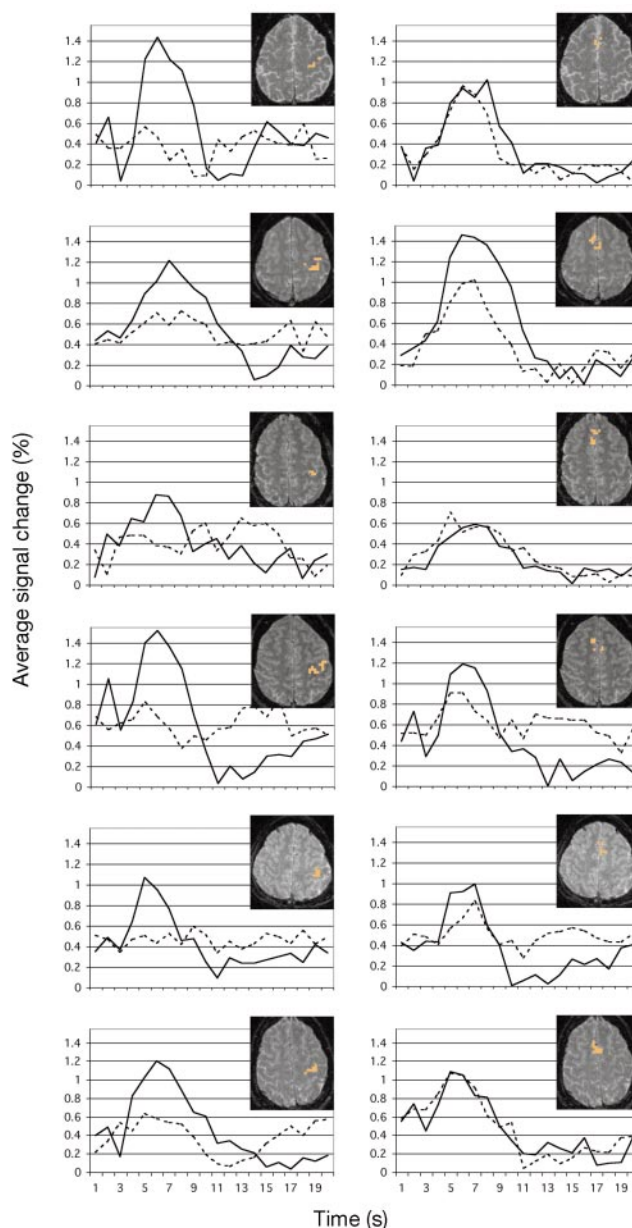


Figure 2 Haemodynamic response curves. These curves were produced by averaging all pixels that passed the threshold of being significantly activated in either the go or the no-go condition. The left column shows the individual results of the primary motor area, the right column of the preSMA. The solid line represents the go-trials, the dash line the no-go trials. The x axes shows the time after the signal, the y axis the averaged signal change. Inset, the pixels that were activated in the motor cortex (left) and the preSMA (right) at the level of the 'hand-knob' (the small posterior protuberance of the pre-central gyrus where the hand is represented). Note that activation extends into neighbouring levels.

impossible²⁵. However, the semi-quantitative statement—that there is a difference in the metabolic demand between inhibition, which does not lead to an increased BOLD signal, and excitation, which leads to an increased BOLD signal—is reliable. We note that with inhibition the frequently encountered phenomenon of ‘deactivation’ (a drop of the magnetic resonance signal intensity below baseline) is not seen, as was suggested correctly on theoretical grounds²⁶. Our data imply that when blood flow is increased, it is very likely that this represents predominant excitatory synaptic activity. This reduces ambiguity in interpreting functional imaging studies.

Could the difference between the two conditions be because the activation we see with the go task is due to afferent input of the moving finger? Two recent positron emission tomography studies help to answer this question. Weeks and co-workers²⁷ found normal activation of the primary sensorimotor area in deafferented individuals performing finger movements, while Mima and co-workers²⁸ were unable to find statistically significant activation of the motor cortex in a carefully controlled passive finger movement task. Therefore, it seems unlikely that the difference we observed is mainly due to afferent input.

We now consider if the no-go task might be showing no BOLD effect because inhibition for this task is shorter than excitation for the go task. From our TMS studies we know that inhibition of the primary motor cortex lasted, on average, as long as the period when TMS was applied (300 ms), which corresponds to the length of a burst pattern during a brief button press in the go task²⁹. Therefore, duration of inhibition is very similar to duration of excitation. Additional information about the time necessary to evoke a haemodynamic response can be gained from examining the preSMA. Single-cell recordings in monkeys showed that most preSMA neurons stopped firing before movement onset¹⁷. Movement onset in our volunteers occurred on average 268 ms after the instruction cue. Thus, cells in the preSMA could not have increased their firing rate for more than about 150 ms, which is shorter than the duration of inhibition in the primary motor cortex. Therefore, duration is unlikely to be too short to evoke a haemodynamic response.

We also need to consider whether the difference in the BOLD signal could be explained by the assumption that more neurons are excited with the go task than are inhibited in the no-go task. We expect that the same neurons that are involved in executing the motor command are inhibited by the ‘non-execution’ of that particular task. However, even if equal numbers of neurons are involved in both conditions, the number of synapses becoming active during inhibition might be smaller, or, alternatively, the inhibitory synapses might be less active than the excitatory synapses. Both possibilities imply that inhibition is more effective than excitation, and therefore support our finding that inhibition is less metabolically demanding than excitation. The same applies naturally if indeed fewer neurons were involved in the inhibitory condition, meaning that in order to be functionally relevant only a little inhibition is necessary.

A further explanation for our findings might be that even if the afferent excitatory and inhibitory input targets the same set of neurons, the excitatory signal may continue to propagate within the intrinsic circuitry of the motor cortex. These population dynamics—induced only by the excitatory afferents—may then account for the difference in the BOLD signal. □

Methods

Single-pulse TMS experiment

Eight healthy right-handed volunteers (ages 21–55 yr) were screened. All had given written informed consent. The study was approved by the local ethical committee. Silver-silver chloride surface electromyography (EMG) electrodes were placed over the flexor digitorum communis and extensor digitorum communis muscles of the dominant arm. The EMG activity was amplified using a conventional EMG machine (Counterpoint Electromyograph, Dantec Electronics) with bandpass between 100 and 2,500 Hz. The signal was digitized at 5 kHz and fed into an IBM-compatible 486AT computer for off-line

analysis. Focal TMS was performed with the target muscles at complete rest (EMG activity < 25 μ V). The motor cortex was excited with a ‘figure-8-shaped’ coil connected to a magnetic stimulator (Magstim 200). With a slightly suprathreshold stimulus intensity, the optimal position for eliciting maximal amplitude MEPs in the flexor digitorum communis muscle was determined. The optimal position was marked to ensure constant coil placement throughout the experiment. As measures of motor excitability, resting motor threshold (rMT) and MEP size were determined. The individual rMT was calculated for the flexor digitorum communis muscle at the optimal stimulation site to the nearest 1% of the maximum stimulator output, and was defined as the minimal stimulus intensity required to produce MEPs of > 50 μ V in at least 5 out of 10 consecutive trials. The individual rMT was expressed relative to the maximal stimulator output. MEP size was determined by averaging the peak-to-peak amplitudes over 10 single trials for each condition (less if the volunteers made a mistake) at a stimulus intensity of 10% of maximum stimulator output above the individual rMT. The time interval between two consecutive trials randomly changed between 17 and 23 s. Go and no-go trials were randomly intermixed. Volunteers sat looking at a screen and were instructed to push a button with their right index finger as quickly as possible after a visual go instruction, but to take no action after a no instruction. TMS was applied 200, 300, 350, 400 and 500 ms after the instruction. Additionally, 20 MEPs were recorded before and after the principal experiment (‘baseline MEPs’), and averages were calculated for comparison with mean MEP obtained during the no-go condition. The MEPs of the no-go trials were significantly smaller than the baseline MEPs in six volunteers ($P < 0.05$, Mann–Whitney U test). These volunteers were selected for the fMRI experiment.

Double-pulse TMS experiment

Intracortical inhibition was tested in four right-handed healthy volunteers (age 21–55 yr) using the paired-pulse TMS protocol¹². The experimental set-up was identical to the one described above with the following exceptions: TMS was always applied 200 ms after the instruction, because this delay consistently showed the strongest MEP inhibition in the single-pulse TMS experiment (Fig. 1). TMS consisted either of a single test pulse or of the test pulse preceded by a conditioning pulse at a 3-ms interstimulus interval. The intensity of the test pulse was adjusted to produce a test MEP of approximately 1 mV in peak-to-peak amplitude when given alone. The intensity of the conditioning pulse was below motor threshold, and adjusted to produce a weak (to ~80%) inhibition of the test MEP at baseline. For each condition, at least ten conditioned and test trials were recorded. Technically appropriate recordings could be obtained in two volunteers (average drift for both the unconditioned and conditioned MEPs in the two baseline conditions $\leq \pm 20\%$). In both volunteers, the previous finding of a strong inhibition in the no-go condition was reproduced in the unconditioned trials (49% and 60% reduction of the MEPs, $P < 0.01$ Mann–Whitney U test). In the conditioned trials, both volunteers showed a further increase in inhibition (32% and 18% decrease of the MEPs in the conditioned compared with the unconditioned trials, $P < 0.05$ Wilcoxon signed rank test).

fMRI experiment

Six volunteers showing inhibition of the motor cortex during the no condition of experiment one were scanned. 98 go/no-go instructions were shown on a screen that could be seen through the mirror of the head coil. The instructions were randomly presented between 17 and 23 s. Seven consecutive scans were taken, each lasting 304 s. The images were realigned using an in-house algorithm³⁰ and analysed using interactive data language (IDL) (RSI Inc.). A multi-linear regression analysis modelled to the expected haemodynamic response curve (a truncated gaussian function with a standard deviation (σ) of 3.5 s and a delay from stimulus to the top of the response of 5 s) was used to select the voxels that passed the threshold for significant activation after Bonferroni correction for the preselected regions of interest (motor cortex and preSMA). The haemodynamic response curve for all the voxels that passed the threshold in either the go or the no-go condition was then calculated by averaging the raw signal from the selected voxels for a period of 20 s after each instruction (Fig. 2).

Received 31 March; accepted 25 May 2000.

- Sokoloff, L. *et al.* The [¹⁴C]deoxyglucose method for the measurement of local cerebral glucose utilization: theory, procedure, and normal values in the conscious and anesthetized albino rat. *J. Neurochem.* **28**, 897–916 (1977).
- Ackermann, R. F., Finch, D. M., Babb, T. L. & Engel, J. Jr Increased glucose metabolism during long-duration recurrent inhibition of hippocampal pyramidal cells. *J. Neurosci.* **4**, 251–264 (1984).
- Nudo, R. J. & Masterton, R. B. Stimulation-induced [¹⁴C]2-deoxyglucose labeling of synaptic activity in the central auditory system. *J. Comp. Neurol.* **245**, 553–565 (1986).
- Rothman, D. L. *et al.* In vivo nuclear magnetic resonance spectroscopy studies of the relationship between the glutamate–glutamine neurotransmitter cycle and functional neuroenergetics. *Phil. Trans. R. Soc. Lond. B* **354**, 1165–1177 (1999).
- Fairen, J. D. & Regador, J. In *Cerebral Cortex* (eds Peters, A. & Jones, E.) 201–253 (Plenum, New York, 1984).
- DeFelipe, J. & Farinas, I. The pyramidal neuron of the cerebral cortex: morphological and chemical characteristics of the synaptic inputs. *Prog. Neurobiol.* **39**, 563–607 (1992).
- Beaulieu, C. & Colonnier, M. A laminar analysis of the number of round-asymmetrical and flat-symmetrical synapses on spines, dendritic trunks, and cell bodies in area 17 of the cat. *J. Comp. Neurol.* **231**, 180–189 (1985).
- Peters, A., Oalaly, S. & Webster, H. *The Fine Structure of the Nervous System. Neurons and their Supporting Cells* (Oxford Univ. Press, New York, 1991).
- Koos, T. & Tepper, J. M. Inhibitory control of neostriatal projection neurons by GABAergic interneurons. *Nature Neurosci.* **2**, 467–472 (1999).
- Buckner, R. L. *et al.* Detection of cortical activation during averaged single trials of a cognitive task using functional magnetic resonance imaging. *Proc. Natl Acad. Sci. USA* **93**, 14878–14883 (1996).

11. Leocani, L., Cohen, L., Wassermann, E., Ikome, K. & Hallett, M. Human corticospinal excitability evaluated with transcranial magnetic stimulation during different reaction time paradigms. *Brain* **123**, 1161–1173 (2000).
12. Rosler, K. M., Etter, C., Truffert, A., Hess, C. W. & Magistris, M. R. Rapid cortical motor output map changes assessed by the triple stimulation technique. *NeuroReport* **10**, 579–583 (1999).
13. Kujirai, T. *et al.* Corticocortical inhibition in human motor cortex. *J. Physiol. (Lond.)* **471**, 501–519 (1993).
14. Ziemann, U., Rothwell, J. C. & Ridding, M. C. Interaction between intracortical inhibition and facilitation in human motor cortex. *J. Physiol. (Lond.)* **496**, 873–881 (1996).
15. van Gelderen, P. *et al.* Three-dimensional functional magnetic resonance imaging of human brain on a clinical 1.5-T scanner. *Proc. Natl Acad. Sci. USA* **92**, 6906–6910 (1995).
16. Deiber, M. P. *et al.* Cortical areas and the selection of movement: a study with positron emission tomography. *Exp. Brain Res.* **84**, 393–402 (1991).
17. Shima, K., Mushiake, H., Saito, N. & Tanji, J. Role for cells in the presupplementary motor area in updating motor plans. *Proc. Natl Acad. Sci. USA* **93**, 8694–8698 (1996).
18. Picard, N. & Strick, P. L. Motor areas of the medial wall: a review of their location and functional activation. *Cereb. Cortex* **6**, 342–353 (1996).
19. Humberstone, M. *et al.* Functional magnetic resonance imaging of single motor events reveals human presupplementary motor area. *Ann. Neurol.* **42**, 632–637 (1997).
20. Frackowiak, R. S., Lenzi, G. L., Jones, T. & Heather, J. D. Quantitative measurement of regional cerebral blood flow and oxygen metabolism in man using ^{15}O and positron emission tomography: theory, procedure, and normal values. *J. Comput. Assist. Tomogr.* **4**, 727–736 (1980).
21. Raichle, M. E. Measurement of local cerebral blood flow and metabolism in man with positron emission tomography. *Federation Proceedings* **40**, 2331–2334 (1981).
22. Fox, P. T. & Raichle, M. E. Focal physiological uncoupling of cerebral blood flow and oxidative metabolism during somatosensory stimulation in human subjects. *Proc. Natl Acad. Sci. USA* **83**, 1140–1144 (1986).
23. Ogawa, S. *et al.* Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proc. Natl Acad. Sci. USA* **89**, 5951–5955 (1992).
24. Silva, A. C., Lee, S. P., Yang, G., Iadecola, C. & Kim, S. G. Simultaneous blood oxygenation level-dependent and cerebral blood flow functional magnetic resonance imaging during forepaw stimulation in the rat. *J. Cereb. Blood Flow Metab.* **19**, 871–879 (1999).
25. Rees, G. *et al.* Characterizing the relationship between BOLD contrast and regional cerebral blood flow measurements by varying the stimulus presentation rate. *Neuroimage* **6**, 270–278 (1997).
26. Raichle, M. E. Behind the scenes of functional brain imaging: a historical and physiological perspective. *Proc. Natl Acad. Sci. USA* **95**, 765–772 (1998).
27. Weeks, R. A., Gerloff, C., Dalakas, M. & Hallett, M. PET study of visually and non-visually guided finger movements in patients with severe pan-sensory neuropathies and healthy controls. *Exp. Brain Res.* **128**, 291–302 (1999).
28. Mima, T. *et al.* Brain structures related to active and passive finger movements in man. *Brain* **122**, 1989–1997 (1999).
29. Hallett, M., Shahani, B. T. & Young, R. R. EMG analysis of stereotyped voluntary movements in man. *J. Neurol. Neurosurg. Psychiatry* **38**, 1154–1162 (1975).
30. Thavenaz, P., Ruttimann, U. E. & Unser, M. Iterative multi-scale registration without landmarks. *Proc. IEEE Int. Conf. on Image Processing* Vol. III 228–231 (IEEE Computer Society Press, Los Alamitos, CA, 1995).

Acknowledgements

We thank D.G. Schoenberg for help with the manuscript.

Correspondence and requests for materials should be addressed to M.H. (e-mail: hallettm@ninds.nih.gov).

A role for the C3a anaphylatoxin receptor in the effector phase of asthma

Alison A. Humbles*, **Bao Lu***, **Christy A. Nilsson***, **Craig Lilly†**, **Elliot Israel†**, **Yuko Fujiwara‡**, **Norma P. Gerard*** & **Craig Gerard***

* Ina Sue Perlmutter Laboratory, Children's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA

† Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA

‡ Division of Hematology and Oncology, Children's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA

Asthma is a chronic inflammatory disease of the airways and lung mucosa with a strong correlation to atopy and acquired (IgE) immunity¹. However, many features of bronchial asthma, such as smooth muscle contraction, mucus secretion and recruitment of inflammatory cells, are consistent with the actions of complement anaphylatoxins, in particular C3a and C5a². Complement

activation forms a central core of innate immune defence against mucosal bacteria, viruses, fungi, helminths and other pathogens. As a system of 'pattern-recognition molecules', foreign surface antigens and immune complexes lead to a proteolytic cascade culminating in a lytic membrane attack^{2,3}. The anaphylatoxins C3a and C5a are liberated as activation byproducts and are potent pro-inflammatory mediators that bind to specific cell surface receptors and cause leukocyte activation, smooth muscle contraction and vascular permeability². Here we show that in a murine model of allergic airway disease, genetic deletion of the C3a receptor protects against the changes in lung physiology seen after allergen challenge. Furthermore, human asthmatics develop significant levels of ligand C3a following intra-pulmonary deposition of allergen, but not saline. We propose that, in addition to acquired immune responses, the innate immune system and complement (C3a in particular) are involved in the pathogenesis of asthma.

To investigate the role of the anaphylatoxins in host defence and inflammation, we have generated mouse strains deficient in the receptors for C5a and C3a. Mice lacking the C5a receptor have altered lung mucosal defence against *Pseudomonas pneumonia*⁴, and have tissue-variable protection against injury mediated by immune complexes⁵. Additionally, evidence for a 'protective' role of C5a in models of brain⁶ and pancreatic inflammation⁷ have been observed using these mice, indicating that the C5a receptor may be able to modulate both pro- and anti-inflammatory effects.

The C3a receptor (C3aR) is an unusual member of the peptidergic G-protein-coupled receptor family: it has a large extracellular loop (172 amino acids) between transmembrane segments 4 and 5 (refs 8, 9). Although the bulk of this loop may be mutated without affecting C3a function at the receptor, the conservation of this unusual feature across species suggests the potential for an alternative ligand¹⁰. Mice have been generated that are deficient in the ligand precursor C3, and thus lack C3a, but these mice are also deficient in the other C3-derived biological activities (for example, phagocytosis and regulation of B-cell function of T-cell-dependent antigens independent of the C3a receptor)¹¹. We generated C3aR homozygous knockout mice on a BALB/c and 129/sv mixed background by homologous recombination with standard techniques. Figure 1 shows the construct, detection and function of the mutant alleles. C3aR-deficient mice were born at the expected mendelian ratios, showed no overt developmental or morphological abnormalities and were fertile when maintained under specific pathogen-free conditions.

Although the function of C3a in disease is not as well defined as that of C5a, there is evidence that C3a/C3aR is involved in allergic disease: C3a has been shown to activate and be chemotactic for human eosinophils and mast cells^{12–14}, and to contract human lung parenchymal strips¹⁵. Therefore, the presence of this receptor on airway smooth muscle and cells associated with allergic disease prompted us to investigate its role in a mouse model of allergic asthma. We used a standard model involving low-dose immunization with ovalbumin (OA) and aluminium hydroxide followed by serial aerosol challenge on days 21–24 to generate allergic airway disease in mice backcrossed three generations against the BALB/c strain. This strain of mouse is routinely used in this model because of the response seen in eosinophilic airway inflammation, Th-2 cytokine production and airway hyper-responsiveness. At both 6 and 24 h after the last aerosol challenge, both wild-type and C3aR-deficient mice developed a similar lung inflammatory infiltrate as assessed by analysis of total white blood cells in the bronchoalveolar lavage (BAL) (Fig. 2a). Interestingly, there was no difference in eosinophil numbers in either the BAL or lung tissue of wild-type versus C3aR-deficient mice when using this repeated OA-challenge protocol (Fig. 2b). IgE levels were also identical and there were no differences in the BAL fluid levels of interleukin (IL)-4, IL-5 and IL-13 (not shown).

We used whole body plethysmography to measure the physiological response of sensitized and sham-treated mice to aerosolized methacholine. Airway hyper-responsiveness (AHR) was measured 24 h after the last aerosol challenge by recording respiratory pressure curves, and was expressed as enhanced pause (Penh), a calculated value that correlates with measurement of airway resistance, obstruction and intrapleural pressure in the same mouse¹⁶. As shown in Fig. 3, when sham-treated wild-type littermates are compared with sensitized/challenged wild-type littermates, airway hyper-responsiveness to inhaled methacholine is markedly enhanced. In contrast, when C3aR-deficient mice are treated identically, there is no statistically significant change in airway responsiveness to methacholine. Therefore, absence of C3aR confers significant protection against increased AHR. Using the same sensitization and challenge protocol, we assessed AHR to methacholine in anaesthetized, tracheostomized, mechanically ventilated mice. This system allows the simultaneous measurement of pul-

monary conductance (GL) and compliance (C_{dyn}) in response to intravenous infusion of a bronchoconstrictor¹⁷. Concomitantly, using two parameters to assess lung function, we found that again sensitized/challenged C3aR-deficient mice were protected from increased AHR to methacholine (Fig. 4). Interestingly, this effect was more pronounced with respect to pulmonary compliance (Fig. 4, right panel).

Because of previous reports associating complement activation with some patients with exercise-induced asthma¹⁸ and status asthmaticus¹⁹, we also measured C3a levels in lung lavage fluids of

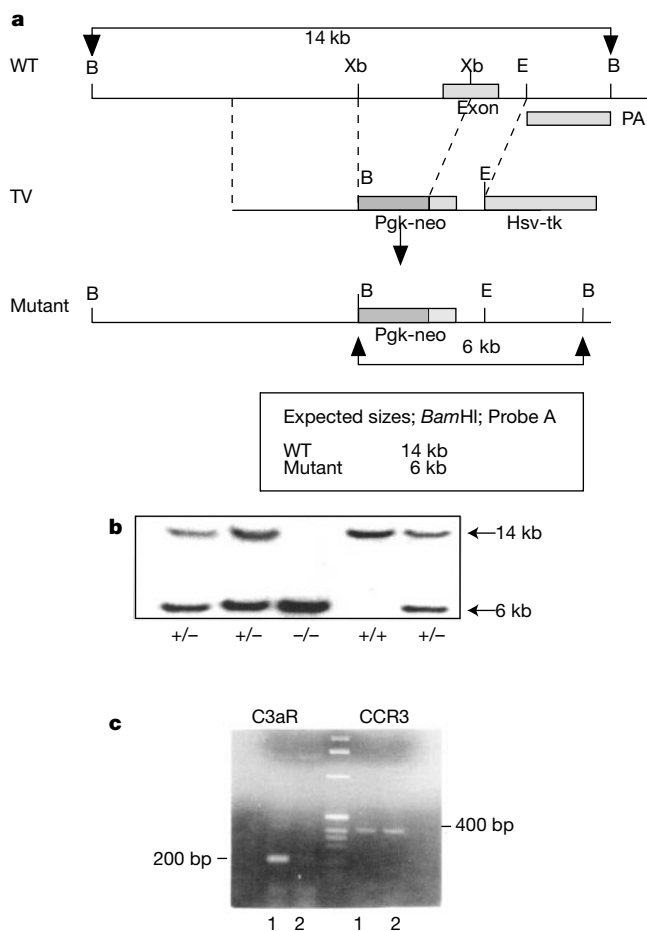


Figure 1 Targeted disruption of the mouse C3aR gene. **a**, Targeting strategy. Wild-type (WT) allele (top), targeting vector (TV; middle) and predicted mutated allele (bottom). The wild-type gene is shown containing a single exon of the receptor as an open block bounded by two *Bam*HI sites. Around 736 base pairs of the N-terminal region including the start codon was deleted and replaced with a neomycin-resistant gene driven by the PGK promoter. Homologous recombination led to a diagnostic pattern on restriction digests probed with a downstream flanking sequence (Probe A/PA). B, *Bam*HI; E, *Eco*RI; Xb, *Xba*I. **b**, Representative Southern blot analysis of tail DNA from littermates. Genomic DNA was digested with *Bam*HI and hybridized with flanking probe A. The 14-kb wild-type and 6-kb targeted allele *Bam*HI fragments identified by probe A are shown. **c**, Polymerase chain reaction with reverse transcriptase (RT-PCR) analysis of C3aR expression using total RNA from bone marrow cells of wild-type and homozygote littermates. RT-PCR was also performed using primers for CCR3 messenger RNA as a control for the presence of amplifiable RNA.

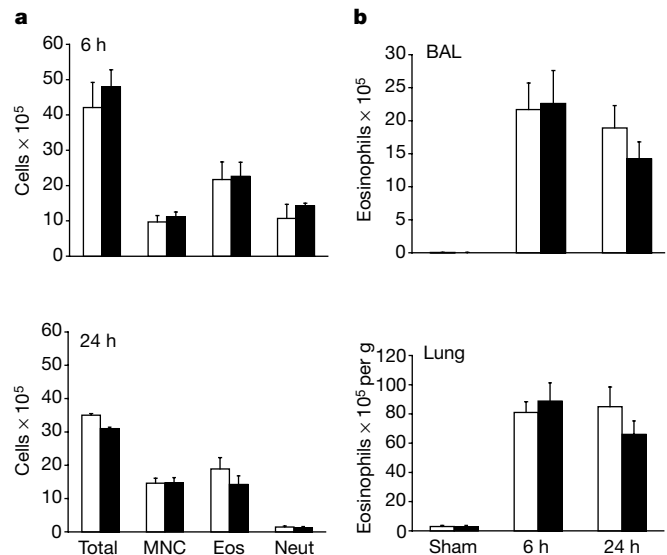


Figure 2 Pulmonary accumulation of inflammatory cells. Sham or sensitized mice were exposed to saline or OA aerosol on days 21–24. **a**, Total cell numbers in BAL from C3aR-deficient (solid bars) or wild-type littermate (open bars) mice were assessed 6 (top) and 24 h (bottom) after the last aerosol challenge. MNC, mononuclear cells; Eos, eosinophils; Neut, neutrophils. **b**, Eosinophil numbers in BAL (top) or lung tissue (bottom) at 6 and 24 h after allergen challenge. Results are presented as means ± s.e.m. (sham, *n* = 8; 6 h, *n* = 10–12 mice per group; 24 h: *n* = 26–27 mice per group).

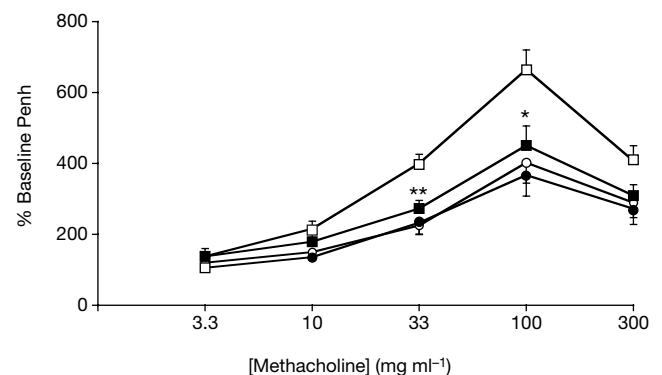


Figure 3 Assessment of AHR to inhaled methacholine (MCh) in conscious mice. Sham-treated wild-type (open circles) or C3aR-deficient (filled circles) mice were exposed to aerosolized saline, and OA-sensitized wild-type (open squares) or C3aR-deficient (filled squares) mice were exposed to aerosolized OA on days 21–24. Airway responses to increasing concentrations of MCh were assessed 24 h after the last aerosol exposure using barometric whole-body plethysmography. The dose–response curves of MCh-induced increases in Penh are shown. Results are expressed as the means ± s.e.m. (sham, *n* = 12 mice per group; sensitized/OA challenged, *n* = 20–22 mice per group) of the per cent increase in Penh compared with the baseline Penh values after saline exposure. Significant differences between sensitized/challenged wild-type and C3aR-deficient mice are indicated as **P* < 0.01 and ***P* < 0.001, as determined by unpaired Student's *t*-test.

asthmatic subjects after segmental allergen challenge. Eight atopic asthmatics underwent bronchoscopy and intrapulmonary challenge with allergen diluent (sham) in one lobe and with allergen, to which they exhibited a positive skin test and a 20% reduction in FEV₁ (the forced expiratory volume in the first second), in the other lobe. Four to eight hours later, at the time of onset of the late-phase response in these subjects, the same lobes were lavaged with saline at repeat bronchoscopy. As an added control, five non-asthmatic adults were lavaged with saline alone. Samples of BAL from all subjects were assayed for C3a/C3a des-Arginine by radioimmunoassay (RIA). As shown in Table 1, we saw a significant increase in C3a anaphylatoxin in the allergen-challenged lobes compared with the sham-treated lobes when the results were expressed as median values with interquartile ranges (IQR; $P < 0.01$). Significantly, as the levels of C3a in normal subjects (non-asthmatics) and sham-challenged asthmatics were identical, it is unlikely that trauma during bronchoscopy induced complement activation, leading to C3 generation. We also found that anaphylatoxin formation in asthmatic subjects was positively correlated with neutrophil numbers in BAL ($R^2 = 0.9$, $P < 0.01$, data not shown).

For the purpose of understanding its mechanism, asthma may be divided into three phases: initiation, propagation and effector. The initiating phase in atopic asthma is believed to involve IgE and mast cells²⁰, whereas this phase in cold-air or exercise-induced asthma is less well understood. In the initiating phase, genetic predisposition to atopy predominates. Polygenic traits involving IL-4, IL-5, IL-13 and the development of IgE antibodies against environmental antigens occur. In the propagation phase, Th-2-polarized T lymphocytes and eosinophils infiltrate the airways, leading to a chronic inflammatory state²¹. In the effector phase, elaboration of spasmogenic substances and secretagogues cause bronchoconstriction and mucus hypersecretion. Effective therapies against the effector phase of asthma have been developed. Arachidonic acid metabolites centred on the cysteinyl leukotrienes are blocked by 5-lipoxygenase inhibitors and leukotriene receptor antagonists, and these drugs are effective in some patients²². However, a significant number of

asthma patients are refractory to these therapies, indicating that parallel pathways may generate other spasmogenic mediators. A number of mediators have been suggested as participating in the effector phase of asthma, including neuropeptides and endothelins^{23,24}. Here, we identify the C3a receptor in the mouse as a significant contributor to the effector phase of allergic bronchoconstriction. We additionally identify the ligand C3a in BAL fluids of allergen- but not sham-challenged human asthmatics during their late-phase response, 4–8 h after allergen delivery to the airway.

What mechanisms might lead to the generation of C3a *in vivo* in atopy and asthma? First, it is possible that in addition to activation of complement through the classical pathways with both IgE and IgG, some macromolecular structures from the dust mite, fungi or ragweed allergen may be recognized by the alternative pathway and fix C3 directly. Pattern recognition of carbohydrate structures by the mannose-binding lectin pathway is also conceivable. However, not all subjects wheeze when exposed to these potentially complement-activating stimuli. Another possibility is that mast cell proteases released directly after IgE engagement of allergen on mast cells/basophils could generate C3a directly, as occurs *in vitro*²⁵. Thus, with C3 levels naturally high in interstitial and lung lining fluids, allergen-triggered release of mast cell proteases in atopic individuals could lead to C3a release without typical activation of the entire complement cascade. Direct anaphylatoxin receptor blockade, which is possible as it is a member of the G-protein-coupled receptor superfamily, might be therapeutic in some human atopic asthma patients. □

Methods

Targeted disruption of the mouse C3aR gene

We screened a mouse 129/sv genomic library with mouse C3aR complementary DNA. An 8-kilobase (kb) genomic fragment containing the C3aR gene was used to construct the targeting vector. We deleted a 3-kb *Xba*I fragment containing a partial amino-terminal and 5' untranslated region (approximately 736 base pairs) and replaced it with a neomycin-resistant gene driven by the PGK promoter. This mutated fragment was subcloned into the pPNT vector for double selection with geneticin and glancyclovir. The targeted vector was linearized and electroporated into J1 embryonic stem cells, and the correctly targeted event was screened by Southern blotting. We injected two targeted cell lines into blastocysts derived from C57BL/6 mice. Chimaeric males were bred with female BALB/c to yield germline transmission of the targeted allele. Heterozygote animals were then bred with wild-type BALB/c mice to give third generation BALB/c backcrossed mice. Heterozygote animals from these mice were then bred against each other to yield homozygote C3aR^{-/-} and C3aR^{+/-} mice, which were used for all studies.

Animal studies

These were performed according to institutional and NIH guidelines for animal use and care.

Immunization and challenge protocol

We immunized mice with 10 µg of OA and 1 mg of aluminium hydroxide, made up in 0.2 ml sterile saline and injected intraperitoneally on days 0, 7 and 14. Sham-immunized mice received aluminium hydroxide alone. On days 21–24, mice were exposed to aerosolized OA (5%) or saline for 40 min. They were killed with a barbiturate overdose 6 or 24 h after the last aerosol challenge and BAL was performed with 2 × 1-ml aliquots of HBSS containing 10 mM HEPES and 10 mM EDTA. We determined total cell numbers in BAL using kimura staining, and performed differential cell counts (of at least 500 cells per slide) on cytospin preparations stained with eosin and methylene blue (Diff-Quik, Dade Diagnostics). Eosinophil numbers in lung tissue were quantified by measuring eosinophil peroxidase as described²⁶.

Determination of AHR in conscious animals

AHR was measured in unrestrained conscious animals using barometric plethysmography¹⁶ (Buxco Electronics). AHR is expressed as the fold increase in Penh (enhanced pause). Twenty-four hours after the last aerosol challenge, mice were placed in whole-body plethysmographic chambers and exposed for 1 min to aerosolized saline and subsequently to increasing concentrations of aerosolized methacholine (3.3–300 mg ml⁻¹). Recordings were taken for 10 min after each aerosolization. The Penh values measured during each 10-min cycle were averaged and expressed for each methacholine concentration as a percentage of the baseline Penh value following saline exposure.

Determination of AHR in anaesthetized animals

Each mouse was anaesthetized, a tracheostomy was created and mechanical ventilation was instituted through a tracheal cannula. We measured the pulmonary mechanical

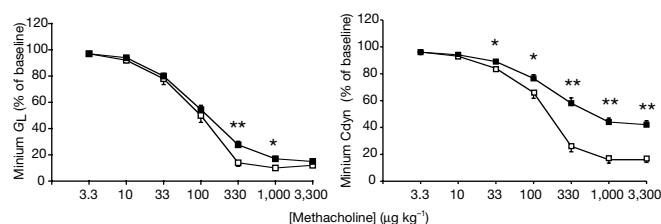


Figure 4 Airway responsiveness to intravenous methacholine (MCh) in anaesthetized mice. OA-sensitized wild-type (open squares) or C3aR-deficient (filled squares) mice were exposed to aerosolized OA on days 21–24. Twenty-four hours after the last aerosol exposure, mice were anaesthetized, intubated and mechanically ventilated, and airway responses to increasing doses of intravenous MCh assessed. The dose–response curves for pulmonary conductance (G_L , left) and pulmonary compliance (C_{dyn} , right) are shown. Results are expressed as the means $\pm$ s.e.m. (wild type, $n = 11$; C3aR^{-/-}, $n = 16$ mice) of the per cent minimal decrease in pulmonary conductance or compliance obtained after MCh challenge compared with the baseline value just before challenge. Significant differences between sensitized/challenged wild-type and C3aR^{-/-} mice are indicated as * $P < 0.04$ –0.01 and ** $P < 0.001$ as determined by unpaired Student's *t*-test.

Table 1 C3a/C3a des-Arginine levels in human BAL fluid

	Normals ($n = 5$)	Asthmatics ($n = 8$)		
		Pre-allergen	Sham	Allergen
Median	59	64	69	123*
IQR	58–107	42–79	46–91	55–357

C3a/C3a des-Arginine levels (expressed in ng ml⁻¹) were measured in BAL fluid by radioimmunoassay. Results shown are the medians and the interquartile ranges (IQR) for five normal and eight asthmatic subjects that underwent segmental sham or allergen challenge.

* $P < 0.01$, significant difference between the sham and allergen groups, as determined by the Wilcoxon signed-ranks test.

parameters GL (pulmonary conductance) and Cdyn (pulmonary compliance) as described¹⁷. Logarithmically increasing doses of methacholine (3.3–3,300 µg kg⁻¹) were administered intravenously through a jugular venous catheter. Maximally reduced GL and Cdyn values after each methacholine dose were expressed as percentages of their values obtained just before the infusion of that dose. Intervals of 3–5 min were allowed to elapse between doses to allow GL and Cdyn to return to within 10% of the baseline obtained before the preceding dose.

Segmental allergen challenge of human subjects

Approval for the human study was granted by the Brigham and Women's Hospital Review Board. The study subjects included five control subjects without any history of respiratory disorder, including asthma, and eight allergic asthmatics who met the American Thoracic Society criteria for an asthma diagnosis²⁷. Each asthmatic subject had an abnormal allergen skin prick test defined as a wheal of greater than 5 mm to 1 of 12 standard allergens in 50% vol/vol glycerol solution (Eastern 10-tree mix, ragweed mix, 3-weed mix, dog epithelium/mixed breeds, timothy grass, *Dermatophagoides farinae*, *D. pteronyssinus* *Alternaria tenuis*, cockroach, *Cladosporium herbarum* and *Aspergillus* mix (all from Greer Laboratories), and cat hair standardized extract (ALK Laboratories)).

Before the study, the subjects were exposed to increasing concentrations of nebulized allergen to determine the concentration required to reduce their FEV₁ by 20%. We performed segmental allergen challenge with seasonal allergens at times when these agents were not in season. At this time all subjects were free of respiratory symptoms and had stable lung function for at least 4 weeks following the inhalation challenge before segmental challenge, for which procedure subjects underwent two serial bronchoscopic procedures 4–8 h apart.

Bronchoscopy was performed under midazolam/fentanyl conscious sedation. For normal and asthmatic pre-allergen samples, the bronchoscope was advanced to airway occlusion in the anterior segment of the left upper lobe and 3 × 50-ml aliquots of warmed sterile saline instilled and gently aspirated; in asthmatics only, the bronchoscope was then moved and advanced into the right upper lobe, where 2 ml of allergen diluent was delivered; after 5 min the bronchoscope was moved again and advanced into the right middle lobe, where 2 ml of allergen solution was delivered. The bronchoscope was then removed, and focal wheezing was confirmed to be present exclusively over the right middle lobe in all subjects by auscultation. Four to eight hours later, a second bronchoscopy was performed to obtain BAL fluid from the right upper lobe (sham sample) and the right middle lobe (allergen sample).

Measurement of C3a/C3a des-Arginine levels in BAL fluid

BAL fluid samples were collected onto ice, strained through a gauze mesh and centrifuged (700g at 4 °C) and stored at –80 °C until assayed. C3a/C3a des-Arginine levels were measured using an RIA kit (Amersham Pharmacia Biotech). Interference from the precursor C3 in BAL samples was avoided by the inclusion of a selective precipitation step before the assay. All samples were assayed undiluted according to the protocol outlined by the manufacturers.

Received 17 May; accepted 18 July 2000.

- Burrows, B., Martinez, F. D., Halonen, M., Barbee, R. A. & Cline, M. G. Association of asthma with serum IgE levels and skin-test reactivity to allergens. *N. Engl. J. Med.* **320**, 271–277 (1989).
- Gerard, N. P. in *Asthma* (eds Barnes, P. J., Grunstein, M. M., Leff, A. R. & Woolcock, A. J.) 639–651 (Lippincott-Raven, Philadelphia, 1997).
- Medzhitov, R. & Janeway, C. A. Jr. Innate immunity: The virtues of a nonclonal system of recognition. *Cell* **91**, 295–298 (1997).
- Höpfen, U. E., Lu, B., Gerard, N. P. & Gerard, C. The C5a chemoattractant receptor mediates mucosal defence to infection. *Nature* **383**, 86–89 (1996).
- Höpfen, U. E., Lu, B., Gerard, N. P. & Gerard, C. Impaired inflammatory responses in the reverse Arthus Reaction through genetic deletion of the C5a receptor. *J. Exp. Med.* **186**, 749–756 (1997).
- Osaka, H. *et al.* Expression of C5a receptor in mouse brain: role in signal transduction and neurodegeneration. *Neuroscience* **88**, 1073–1082 (1999).
- Bhatia, M. Complement factor C5a exerts an anti-inflammatory effect in acute pancreatitis and pancreatitis-associated lung injury. *Am. J. Physiol.* (submitted).
- Tornetta, M. A., Foley, J. J., Sarau, H. M. & Ames, R. S. The mouse anaphylatoxin C3a receptor. *J. Immunol.* **158**, 5277–5282 (1997).
- Hsu, M. H. *et al.* Cloning and functional characterization of the mouse C3a anaphylatoxin receptor gene. *Immunogenetics* **47**, 64–72 (1997).
- Chao, T. H. *et al.* Role of the second extracellular loop of human C3a receptor in agonist binding and receptor function. *J. Biol. Chem.* **274**, 9721–9728 (1999).
- Fischer, W. H. *et al.* Regulation of the B cell response to T-dependent antigens by classical pathway complement. *J. Immunol.* **157**, 549–556 (1996).
- Daffern, P. J. *et al.* C3a is a chemotaxin for human eosinophils but not for neutrophils. I. C3a stimulation of neutrophils is secondary to eosinophil activation. *J. Exp. Med.* **181**, 2119–2127 (1995).
- Nilsson, G. *et al.* C3a and C5a are chemotaxins for human mast cells and act through distinct receptors via a pertussis toxin-sensitive signal transduction pathway. *J. Immunol.* **157**, 1693–1698 (1996).
- Stimler-Gerard, N. P. & Galli, S. J. Mast cells in anaphylatoxin-induced spasmogenic activity. *J. Immunol.* **138**, 1908–1913 (1987).
- Stimler, N. P., Gerard, C. & O'Flaherty, J. T. in *Platelet Activating Factor* (eds Benveniste, J. & Arnoux, B.) 195–203 (Elsevier, Amsterdam, 1983).
- Hamelmann, E. *et al.* Noninvasive measurement of airway responsiveness in allergic mice using barometric plethysmography. *Am. J. Respir. Crit. Care Med.* **156**, 766–775 (1992).
- Martin, T. R., Gerard, N. P., Galli, S. J. & Drazen, J. M. Pulmonary responses to bronchoconstrictor agonists in the mouse. *J. Appl. Physiol.* **64**, 2318–2323 (1988).
- Smith, J. K., Chi, D. S., Krish, G., Reynolds, S. & Cambron, G. Effect of exercise on complement activity. *Ann. Allergy* **65**, 304–310 (1990).

- Onodera, H. *et al.* Complement system in status asthmatics: analysis of anti-complementary effects induced by methylprednisolone. *Arerugi* **41**, 1567–1574 (1997).
- Pepys, J. & Hutchcroft, B. J. Bronchial provocation tests in etiologic diagnosis and analysis of asthma. *Am. J. Respir. Crit. Care Med.* **112**, 829–859 (1975).
- Bradley, B. L. *et al.* Eosinophils, T-lymphocytes, mast cells, neutrophils and macrophages in bronchial biopsy specimens from atopic subjects with asthma: comparison with biopsy specimens from atopic subjects without asthma and normal control subjects and relationship to bronchial hyperresponsiveness. *J. Allergy Clin. Immunol.* **88**, 661–674 (1991).
- Drazen, J. M. Asthma therapy with agents preventing leukotriene synthesis or action. *Proc. Assoc. Am. Physicians* **111**, 547–559 (1999).
- Barnes, P. J. & Belvisi, M. G. in *Asthma* (eds Barnes, P. J., Grunstein, M. M., Leff, A. R. & Woolcock, A. J.) 1051–1063 (Lippincott-Raven, Philadelphia, 1997).
- Mattoli, S., Soloperto, M., Marini, M. & Fasoli, A. Levels of endothelin in the bronchoalveolar lavage fluid of patients with symptomatic asthma and reversible airflow obstruction. *J. Allergy Clin. Immunol.* **88**, 376–384 (1991).
- Schwartz, L. B. *et al.* Generation of C3a anaphylatoxin from human C3 by mast cell tryptase. *J. Immunol.* **130**, 1891–1895 (1983).
- Humbles, A. A. *et al.* Kinetics of eotaxin generation and its relationship to eosinophil accumulation in allergic airways disease: analysis in a guinea pig model *in vivo*. *J. Exp. Med.* **186**, 601–612 (1997).
- American Thoracic Society. Lung function testing: selection of reference values and interpretative strategies. *Am. Rev. Respir. Dis.* **144**, 1202–1218 (1991).

Acknowledgements

We thank T. Martin and J. Drazen for comments and advice; J. Brewer for assistance; D. Zurakowski for help and advice with analysing the statistics; and the staff of ARCH at Children's Hospital for animal care. This work was supported in part by grants from the NIH to N.P.G. and C.G. at the Perlmutter laboratory and Pfizer Central Research.

Correspondence and requests for materials should be addressed to C.G. (e-mail: gerard_c@gonzo.tch.harvard.edu).

A voltage-dependent channel involved in nutrient uptake by red blood cells infected with the malaria parasite

Sanjay A. Desai^{*†}, Sergey M. Bezrukov[‡] & Joshua Zimmerberg[†]

^{*} The Laboratory of Parasitic Diseases, National Institute of Allergy and Infectious Diseases, [‡] The Laboratory of Physical and Structural Biology, and [†] The Laboratory of Cellular and Molecular Biophysics, National Institute of Child Health and Development, National Institutes of Health, Bethesda, Maryland 20892, USA

Growth of the malaria parasite in human red blood cells (RBCs) is accompanied by an increased uptake of many solutes including anions¹, sugars², purines³, amino acids⁴ and organic cations⁵. Although the pharmacological properties and selectivity of this uptake suggest that a chloride channel is involved, the precise mechanism has not been identified. Moreover, the location of this uptake in the infected RBC is unknown because tracer studies are complicated by possible uptake through fluid-phase pinocytosis⁶ or membranous ducts⁷. Here we have studied the permeability of infected RBCs using the whole-cell voltage-clamp method. With this method, uninfected RBCs had ohmic whole-cell conductances of less than 100 pS, consistent with their low tracer permeabilities⁸. In contrast, trophozoite-infected RBCs exhibited voltage-dependent, non-saturating currents that were 150-fold larger, predominantly carried by anions and abruptly abolished by channel blockers. Patch-clamp measurements and spectral analysis confirmed that a small (< 10 pS) ion channel on the infected RBC surface, present at about 1,000 copies per cell, is responsible for these currents. Because its pharmacological properties and substrate selectivities match those seen with tracer studies, this channel accounts for the increased uptake of small solutes in infected RBCs. The surface location of this new channel

and its permeability to organic solutes needed for parasite growth indicate that it may have a primary role in a sequential diffusive pathway for parasite nutrient acquisition.

Red blood cells have been difficult to patch clamp because of their small size, high deformability and atypical membrane composition. Whole-cell voltage clamp requires the added complexity of rupturing the small patch of membrane under the pipette, and has not been used to study the permeabilities of human RBCs. We have now obtained whole-cell recordings on these cells by matching the oncotic and hydrostatic forces imposed on the cell when the patch is ruptured with brief electrical pulses (400–600 mV, 0.5–1 ms long). Uninfected human RBCs have small ohmic whole-cell conductances (< 100 pS, $n = 12$ cells; Fig. 1a) in this configuration. We used this method to study ion permeation in single RBCs infected with the human malaria parasite, *Plasmodium falciparum*.

Figure 1b shows the whole-cell current in response to voltage steps applied to a trophozoite-infected RBC. This current is 150-fold larger than that of uninfected RBCs ($n = 13$ infected cells; $P < 0.0001$). It exhibited inward rectification (the plateau at positive voltages; Fig. 1c, filled circles) despite using identical solutions on both sides of the RBC membrane, indicating either voltage-dependent gating or rectification of one or more transport pathways in the membrane. We found that the currents could be abruptly abolished by addition of 125 μ M furosemide to the bath solution (Fig. 1c, open circles). Furthermore, these whole-cell currents did not saturate at up to 1.1 M salt solutions (Fig. 1d). These findings suggest a single pathway with channel-type transport on the infected RBC membrane.

To examine the ion selectivity of this pathway, we applied a NaCl gradient across the RBC membrane (Fig. 1e). The reversal potential, E_{rev} , under these conditions shifted to -47 mV (mean of 5 cells), approximating the Nernst potential for Cl^- in this experiment (-49 mV measured with a Cl^- electrode). This E_{rev} indicates a strong selectivity for Cl^- over Na^+ and is consistent with tracer studies. (We calculate a Na^+ to Cl^- permeability ratio, P_{Na}/P_{Cl} , of roughly 10^{-5} from the data in refs 1 and 9.)

The pharmacological properties of this conductive pathway are also consistent with previous tracer studies: furosemide, NPPB (5-nitro-2-(3-phenylpropylamino)benzoic acid), niflumate¹, glybenclamide¹⁰ and phloridzin¹¹ reduced the whole-cell current in low micromolar concentrations, whereas DIDS¹² (4,4'-diisothiocyanatostibene-2,2'-disulphonic acid) had only a modest effect (Table 1). Interestingly, phloridzin had a greater blocking effect in

the pipette solution than in the bath, implicating an intracellular site of action as previously proposed¹¹.

Permeability ratios (P/P_{Cl}) for other anions, determined by E_{rev} measurements, indicate weak selectivity of this pathway, with high permeation rates for large organic anions (Table 2). We found a permeability sequence of $SCN^- > I^- > Br^- > Cl^- > acetate^- > lactate^- > glutamate^-$, corresponding to the Eisenman selectivity sequence number 1 for anions. This suggests permeation through a pathway that has a weak binding site with a large radius¹³; it is inconsistent with a rigid non-interacting pore. (Using data in refs 1, 2 and 9, we calculated P/P_{Cl} for lactate and glutamate to be about 0.5 and 0.03, respectively, which are comparable with our measurements.)

This whole-cell current on infected cells requires transmembrane movement of anions from the bath solution into RBC cytosol. This might occur directly through the exposed RBC plasma membrane. Alternatively, anions might bypass the RBC membrane by moving down the proposed parasitophorous duct (a membranous tubule that may provide the parasite access to serum⁷), and then enter the RBC cytosol through known channels¹⁴ on the parasitophorous vacuolar membrane (PVM), which surrounds the parasite.

To distinguish between surface and duct-associated sites, we used the on-cell patch-clamp method to measure directly transport across patches of RBC plasma membrane. Membrane patches on infected cells often showed a rectifying current (Fig. 2a), suggesting a small conductance ion channel on the infected RBC surface. Prompted by these findings, we used a solution with a high Cl^- concentration (1.15 M) in the bath and pipette to increase conductance. This manoeuvre revealed unambiguous channel activity with fast gating and a slope conductance of 20 pS (Fig. 2b, c), suggesting permeation through a proteinaceous pore rather than a nonspecific membrane defect¹⁵. This channel exhibited voltage-dependent gating which explains the inward rectification seen with whole-cell measurements (Fig. 2d). It was present in high density on trophozoite-infected cells (37 of 46 patches, 19 with > 3 channels), never on uninfected red cells (22 patches) and could be blocked with 200 μ M furosemide in the pipette (much reduced channel activity, $n = 4$ patches on 4 infected cells; Fig. 2e).

This small channel exhibits complex gating with bursts of fast flickering open events and long intervening closed periods (Fig. 2c). Consistent with this complex gating, the spectral density of an on-cell recording revealed a $1/f$ profile¹⁶ (Fig. 3, bottom curve). (A channel with simple, non-bursting gating would have a Lorentzian

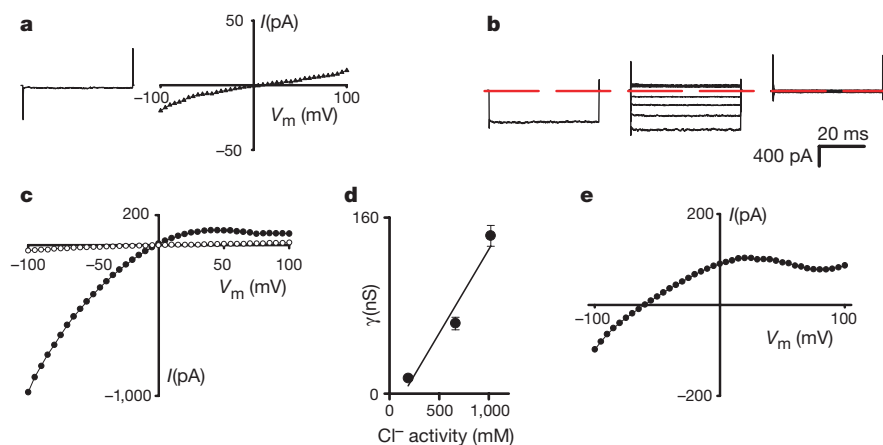


Figure 1 Whole-cell recordings. **a**, Uninfected RBC in solution A. Left, whole-cell current response to a -70 mV pulse. Note the small current amplitude and the capacitive spikes, which reflect patch rupture. Right, corresponding $I-V$ plot. **b**, Trophozoite-infected RBC. Left, recorded as in **a**. Current is 150 times larger than that on uninfected RBCs. Middle and right, responses to pulses of -80 to $+80$ mV (20-mV increments) before and after bath addition of 125 μ M furosemide, respectively. Dashed line represents zero current.

c, Corresponding $I-V$ plots before (filled circles) and after (open circles) addition of furosemide. **d**, Whole-cell chord conductances in solution A plus 0, 500 or 1,000 mM choline $^+$ Cl^- . Chord conductances, determined between -50 and -80 mV, are plotted against Cl^- activity (mean $\pm$ s.e.m. of 12, 7 and 8 cells, respectively). **e**, Whole-cell $I-V$ plot with solution A in the bath and solution B in the pipette. The negative E_{rev} (or x intercept) confirms a predominance of Cl^- permeation.

spectral density.) Spectral analysis of our whole-cell recordings also revealed a $1/f$ profile in 140 mM and 1.15 M Cl^- solutions (Fig. 3, middle and top curves, respectively). These parallel profiles, combined with its voltage-dependent gating and the effects of furosemide (Fig. 2d, e), strongly suggest that the small surface channel is the predominant conductive pathway in the infected RBC membrane. Because the spectra of individual channels add up to the whole-cell spectrum, this method also provides an accurate estimate of the channel's functional copy number on this trophozoite-infected cell at about 1,000 (calculated from the single-channel and whole-cell spectra with 1.15 M Cl^-). Calculations using our single-channel and whole-cell conductances, although less precise than spectral methods, indicate that the functional copy number may rise to about 8,000 per cell depending on parasite maturity.

Because its pharmacology and selectivity match those identified by more than 20 years of tracer flux and osmotic lysis experiments^{1–5,9–12}, this voltage-dependent channel fully accounts for the increased uptake of small solutes into infected RBCs by the “new permeation pathways”. Tracer studies have revealed that this channel is permeable to sugars², purines³, amino acids⁴ and some organic cations such as choline⁵, all of which are required for parasite growth¹⁷. By localizing this channel to the RBC plasma membrane, our findings indicate that it may have a role in acquisition of these nutrients from serum. Our model is a sequential diffusive pathway for nutrient uptake by the parasite (Fig. 4). Small nutrient solutes in serum can enter the RBC cytosol through this channel. (Whereas most nutrients enter predominantly through this channel, sugars enter primarily through the RBC's endogenous carriers¹⁸.) They would then diffuse through the RBC cytosol and cross the PVM through non-selective 150 pS channels¹⁴ (which show different gating from the surface channel) into the vacuolar space. Finally, they would be taken up by the parasite by specific carriers, some

recently identified¹⁹. Our identified channel is probably also involved in metabolic waste removal (such as lactate, Table 2) and volume regulation of infected RBCs¹. Its small single-channel conductance and strong selectivity against Na^+ , despite permeability to bulky uncharged and charged metabolites, pose intriguing pore-design problems, which may be partly solved by its weak binding site (Eisenman selectivity sequence number 1 for anions) and its voltage-dependence.

With all the steps for this sequential nutrient acquisition pathway identified, the need for a parasitophorous duct to provide a parallel route for uptake of serum nutrients becomes less clear. Moreover, the parasitophorous duct—a proposal based on fluorescent macro-molecule uptake by infected cells—remains controversial because

Table 1 Pharmacological properties of the whole-cell currents

Channel blocker	% inhibition
Furosemide	
1 μM (5)	30 $\pm$ 12
15 μM (5)	54 $\pm$ 9
100 μM (11)	83 $\pm$ 2
Glybenclamide	
100 μM (5)	85 $\pm$ 9
Niflumate	
2 μM (3)	20 $\pm$ 9
18 μM (3)	56 $\pm$ 8
78 μM (2)	87
NPPB	
0.8 μM (5)	41 $\pm$ 5
8 μM (4)	78 $\pm$ 8
32 μM (4)	92 $\pm$ 2
DIDS	
100 μM (2)	7
Phloridzin	
300 μM (2)	34
50 μM pipette (3)*	75 $\pm$ 2

The percentage inhibition (mean $\pm$ s.e.m.) of whole-cell currents on bath addition of channel blockers at indicated concentrations.

* Phloridzin was also tested by addition to the pipette and chord conductance measurement as in Fig. 1d.

Table 2 Anion selectivity properties of the whole-cell currents

Anion	E_{rev}	P/P_{Cl}
SCN^- (5)	+50 $\pm$ 3	7.4
I^- in bath (4)*	–31 $\pm$ 5	3.5
Br^- (4)	+4 $\pm$ 1	1.2
Acetate $^-$ (3)	–4 $\pm$ 1	0.87
Lactate $^-$ (4)	–24 $\pm$ 6	0.43
Glutamate $^-$ (5)	–51 $\pm$ 3	0.11

The permeabilities of test anions relative to that of Cl^- .

* Iodide was tested by switching the bath and pipette solutions and using an agar bridge. $\text{Na}_2\text{S}_2\text{O}_3$ (1.1 mM) was added to the NaI solution as an antioxidant. The number of cells tested at each concentration is given in parentheses.

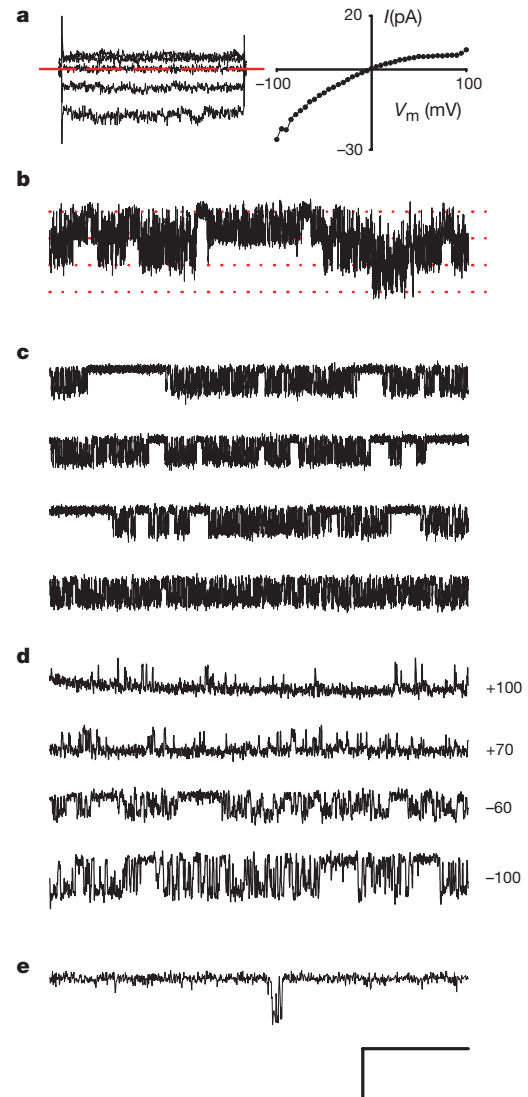


Figure 2 On-cell currents on trophozoite-infected RBCs. **a**, Left, currents in solution A. V_m (membrane potential, negative of pipette potential) pulses were from –80 to +80 mV (40-mV increments). Dashed line represents zero current. This patch had several small channels. Right, corresponding I – V plot. **b**, Currents in solution A plus 1,000 mM choline $^+\text{Cl}^-$. This on-cell patch had three independent and identical channels, whose opening events are represented by downward deflections. V_m , –100 mV. **c**, A single-channel patch showing fast flickering and complex burst behaviour. Seal resistance was 150 G Ω . **d**, Voltage-dependent gating at indicated V_m (millivolts). The channel opens more frequently at negative V_m (downward deflections) than at positive V_m (upward deflections). **e**, Identical conductance but fewer opening events with 200 μM furosemide in the pipette. V_m , –100 mV (Compare the bottom trace in **d**, without furosemide.) Scale bars, 19.5 pA/29 ms (**a**); 4.4 pA/71 ms (**b**); 4.4 pA/48 ms (**c**); 3 pA/12 ms (**d**, **e**).

of poor reproducibility²⁰ and an absence of definitive electron microscopic evidence. If the duct exists, it should produce a furosemide-insensitive current in our whole-cell measurements that involves diffusion down the duct, into the vacuolar space around the parasite and through the PVM channels¹⁴ to the recording pipette. We could not distinguish the currents on furosemide-treated infected cells from those of uninfected cells, implying no electrical continuity between the bath and the vacuole around the parasite. Other models for the proposed duct, such as an intermittently present duct or compartmentalization of the vacuolar space with tight junctions, cannot be excluded with our measurements.

This voltage-dependent channel, found only on infected RBCs, may be either a modified host RBC protein²¹ or a parasite-encoded protein targeted to the host membrane²². In either case, its exposed

location on infected cells and presence in all studied plasmodial species²³ suggest that it may be an ideal target for future chemotherapy development. □

Methods

Whole-cell conductances

We took human RBCs from a healthy volunteer, washed them in solution A (in mM: 115 NaCl, 10 MgCl₂, 5 CaCl₂, 20 Na-HEPES, pH 7.4), and allowed them to settle before patch clamp. Pipettes were pulled from Corning 7056 glass to tip diameters $\leq 0.5 \mu\text{m}$ (1–5 M Ω), heat-polished, filled with filtered pipette solutions and used immediately. For whole-cell recordings, we obtained patch rupture by applying electrical pulses until a capacitive transient developed. Although the whole-cell configuration is usually achieved on other cells by suction applied to the pipette²⁴, this often damaged the small RBC. *Plasmodium falciparum* trophozoites, grown in the same RBCs²⁵, were voltage clamped by the same methods. Whole-cell recordings were filtered at 10 kHz (8 pole Bessel filter). Data for current–voltage profiles represent the current averaged over 20–50 ms. Furosemide had no effect on uninfected RBC currents (data not shown).

Anion selectivity

We determined Na⁺ versus Cl[−] selectivity with a uncharged solute as the primary osmolyte in the pipette (solution B: (in mM) 230 sorbitol, 20 Na-HEPES, 10 MgCl₂, 0.5 EGTA, pH 7.4) to quantitate relative influx of cations and anions. Similar results were obtained with sucrose as the primary osmolyte in the pipette (not shown).

Selectivity of anions was determined under biionic conditions with a bath solution of (in mM) 500 NaCl, 10 MgCl₂, 5 CaCl₂, 20 Na-HEPES, pH 7.4, and pipette solutions of (in mM) 500 Na⁺ salt of the test anion (436 for lactate[−]), 10 MgCl₂, 20 Na-HEPES, pH 7.4. P/P_{Cl} were then calculated with the Goldman equation²⁶ using the mean E_{rev} .

Channel characterization

We carried out single-channel measurements in hypertonic saline to increase conductance through this channel and to reduce pipette noise by reducing access resistance, both of which improved our signal-to-noise ratio. Single-channel recordings were filtered at 6 kHz and digitized at 100 kHz.

We carried out spectral analysis of whole-cell and single-channel recordings on 1-s current sweeps recorded at V_m of +40 mV after filtering at 15 kHz (8 pole Butterworth filter). For each trace, the power spectrum of non-channel noise (for example, from equipment), determined at $V_m = 0$, was subtracted. The power spectrum of a single burst of channel activity was adequately described by a Lorentzian with a characteristic time of 0.5 ms (not shown).

Received 30 May; accepted 2 July 2000.

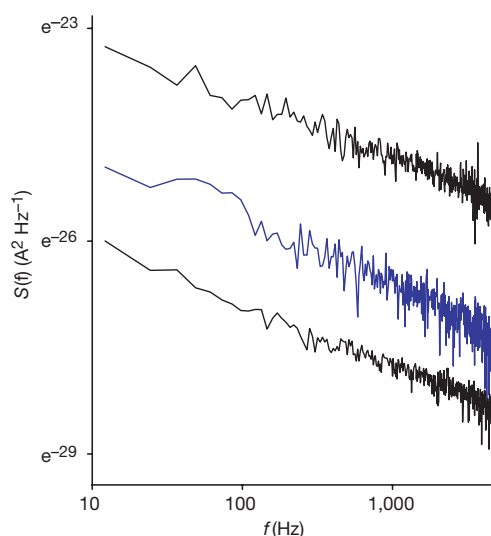


Figure 3 Power spectras of whole-cell (top and middle traces) and single-channel (bottom trace) currents. Spectra were calculated from recordings in solution A (middle trace) or solution A plus 1,000 mM choline⁺Cl[−] (top and bottom traces). These profiles are roughly parallel, indicating that the single-channel type seen with on-cell patch clamp can fully account for the whole-cell currents. The $1/f$ profiles reflect the complexity of single-channel bursting.

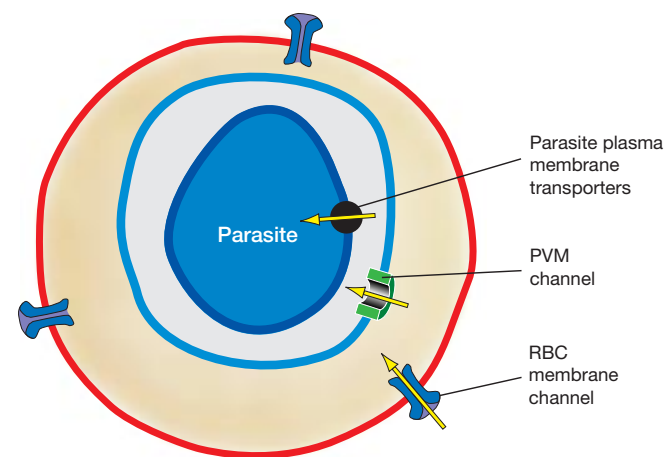


Figure 4 Sequential route of nutrient acquisition. In addition to entry through endogenous RBC carriers (not shown), key nutrients in serum are acquired by sequential passage through the voltage-dependent anion channel on the RBC membrane, the non-selective PVM channel and carriers on the parasite plasma membrane. The parasitophorous duct, if present, would produce continuity between the PVM and RBC membranes. Although not shown here, the PVM has extensive invaginations (called the tubulovesicular network²⁷) that probably increase its surface area for nutrient uptake by the PVM channel.

- Kirk K., Horner, H. A., Elford, B. C., Ellory, J. C. & Newbold, C. I. Transport of diverse substrates into malaria-infected erythrocytes via a pathway showing functional characteristics of a chloride channel. *J. Biol. Chem.* **269**, 3339–3347 (1994).
- Ginsburg, H., Kutner, S., Krugliak, M. & Cabantchik, Z. I. Characterization of permeation pathways appearing in the host membrane of *Plasmodium falciparum* infected red blood cells. *Mol. Biochem. Parasitol.* **14**, 313–322 (1985).
- Upston, J. M. & Gero, A. M. Parasite-induced permeation of nucleosides in *Plasmodium falciparum* malaria. *Biochim. Biophys. Acta* **1236**, 249–258 (1995).
- Cabantchik, Z. I. Properties of permeation pathways induced in the human red cell membrane by malaria parasites. *Blood Cells* **16**, 421–432 (1990).
- Staines, H. M., Rae, C. & Kirk, K. Increased permeability of the malaria-infected erythrocyte to organic cations. *Biochim. Biophys. Acta* **1463**, 88–98 (2000).
- Burns, E. R. & Pollack, S. P. *falciparum* infected erythrocytes are capable of endocytosis. *In Vitro Cell Dev. Biol.* **24**, 481–486 (1988).
- Pouvelle, B. et al. Direct access to serum macromolecules by intraerythrocytic malaria parasites. *Nature* **353**, 73–75 (1991).
- Hunter, M. J. Human erythrocyte anion permeabilities measured under conditions of net charge transfer. *J. Physiol. (Lond.)* **268**, 35–49 (1977).
- Kirk, K. & Horner, H. A. Novel anion dependence of induced cation transport in malaria-infected erythrocytes. *J. Biol. Chem.* **270**, 24270–24275 (1995).
- Kirk, K., Horner, H. A., Spillet, D. J. & Elford, B. C. Glibenclamide and meglitinide block the transport of low molecular weight solutes into malaria-infected erythrocytes. *FEBS Lett.* **323**, 123–128 (1993).
- Kutner, S., Breuer, W. V., Ginsburg, H. & Cabantchik, Z. I. On the mode of action of phlorizin as an antimalarial agent in *in vitro* cultures of *Plasmodium falciparum*. *Biochem. Pharmacol.* **36**, 123–129 (1987).
- Breuer, W. V., Kutner, S., Sylphen, J., Ginsburg, H. & Cabantchik, Z. I. Covalent modification of the permeability pathways induced in the human erythrocyte membrane by the malarial parasite *Plasmodium falciparum*. *J. Cell. Physiol.* **133**, 55–63 (1987).
- Wright, E. M. & Diamond, J. M. Anion selectivity in biological systems. *Physiol. Rev.* **57**, 109–156 (1977).
- Desai, S. A., Krogstad, D. J. & McCleskey, E. W. A nutrient-permeable channel on the intraerythrocytic malaria parasite. *Nature* **362**, 643–646 (1993).
- Ginsburg, H. & Stein, W. D. Biophysical analysis of novel transport pathways induced in red blood cell membranes. *J. Membr. Biol.* **96**, 1–10 (1987).
- Sauve, R. & Szabo, G. Interpretation of $1/f$ fluctuations in ion conducting membranes. *J. Theor. Biol.* **113**, 501–516 (1985).
- Divo, A. A., Geary, T. G., Davis, N. L. & Jensen, J. B. Nutritional requirements of *Plasmodium falciparum* in culture. I. Exogenously supplied dialyzable components necessary for continuous growth. *J. Protozool.* **32**, 59–64 (1985).

18. Kirk, K., Horner, H. A. & Kirk, J. Glucose uptake in *Plasmodium falciparum*-infected erythrocytes is an equilibrative not an active process. *Mol. Biochem. Parasitol.* **25**, 195–205 (1996).
19. Carter, N. S. *et al.* Isolation and functional characterization of the PfNT1 nucleoside transporter gene from *Plasmodium falciparum*. *J. Biol. Chem.* **275**, 10683–10691 (2000).
20. Hibbs, A. R., Stenzel, D. J. & Saul, A. Macromolecular transport in malaria—does the duct exist? *Eur. J. Cell Biol.* **72**, 182–188 (1997).
21. Crandall, I. & Sherman, I. W. *Plasmodium falciparum* (human malaria)-induced modifications in human erythrocyte band 3 protein. *Parasitology* **102**, 335–340 (1991).
22. Bull, P. C. *et al.* Parasite antigens on the infected red cell surface are targets for naturally acquired immunity to malaria. *Nature Med.* **4**, 358–360 (1998).
23. Elford, B. C., Cowan, G. M. & Ferguson, D. J. Parasite-regulated membrane transport processes and metabolic control in malaria-infected erythrocytes. *Biochem. J.* **308**, 361–374 (1995).
24. Hamill, O. P., Marty, A., Neher, E., Sakmann, B. & Sigworth, F. J. Improved patch-clamp techniques for high-resolution current recording from cells and cell-free membrane patches. *Pflügers Arch.* **391**, 85–100 (1981).
25. Trager, W. & Jensen, J. B. Human malaria parasites in continuous culture. *Science* **193**, 673–675 (1976).
26. Goldman, D. E. Potential, impedance, and rectification in membranes. *J. Gen. Physiol.* **27**, 36–60 (1943).
27. Lauer, S. A., Rathod, P. K., Ghori, N. & Halder, K. A membrane network for nutrient import in red cells infected with the malaria parasite. *Science* **276**, 1122–1125 (1997).

Acknowledgements

We thank L. Miller for steady help and advice, and S. Brazer, F. Cohen, A. Sher and T. Welles for reviewing the manuscript.

Correspondence and requests for materials should be addressed to S.A.D. (e-mail sdesai@niaid.nih.gov).

Effects of oncogenic mutations in *Smoothed* and *Patched* can be reversed by cyclopamine

Jussi Taipale*, James K. Chen*, Michael K. Cooper†*, Baolin Wang*, Randall K. Mann*, Ljiljana Milenkovic‡, Matthew P. Scott‡ & Philip A. Beachy*

* Department of Molecular Biology and Genetics, and † Department of Neurology, Howard Hughes Medical Institute, The Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

‡ Departments of Developmental Biology and Genetics, Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California 94305-5427, USA

Basal cell carcinoma, medulloblastoma, rhabdomyosarcoma and other human tumours are associated with mutations that activate the proto-oncogene *Smoothed* (*SMO*) or that inactivate the tumour suppressor *Patched* (*PTCH*). *Smoothed* and *Patched* mediate the cellular response to the Hedgehog (Hh) secreted protein signal, and oncogenic mutations affecting these proteins cause excess activity of the Hh response pathway^{1,2}. Here we show that the plant-derived teratogen cyclopamine, which inhibits the Hh response^{3,4}, is a potential 'mechanism-based' therapeutic agent for treatment of these tumours. We show that cyclopamine or synthetic derivatives with improved potency block activation of the Hh response pathway and abnormal cell growth associated with both types of oncogenic mutation. Our results also indicate that cyclopamine may act by influencing the balance between active and inactive forms of *Smoothed*.

Whereas embryonic loss of Sonic hedgehog (Shh) signalling can result in cyclopia and other developmental defects⁵, inappropriate activation of the Shh response pathway is associated with several types of human tumour^{6–13}. Current approaches to treatment of such neoplastic disorders are limited by the cytotoxic effects of therapeutic agents on proliferating tissues. Alternative 'mechanism-based' approaches specifically targeting abnormally active signalling pathways in defined types of cancer¹⁴ might avoid such toxicity,

particularly if the pathways in question functioned primarily in embryonic development and were not required for survival in adults. Cyclopamine, a plant steroidal alkaloid, induces cyclopia in vertebrate embryos¹⁵ and has been shown to act by inhibiting the cellular response to the Shh signal^{3,4}. To evaluate the therapeutic potential of cyclopamine for the treatment of Hh-pathway-associated disorders we investigated the mechanism by which cyclopamine acts.

Cellular responses to the Hh signal are controlled by two transmembrane proteins, Smo and Ptch, which are predicted to have seven and twelve transmembrane spans, respectively. Genetic and biochemical evidence indicates that Ptch suppresses the activity of Smo, and that binding of Hh to Ptch relieves this suppression^{1,2}, allowing activation of downstream targets through the Ci/Gli family of transcriptional effectors^{16–19}. As our previously described Hh signalling assay used *Drosophila melanogaster* cultured cells and was not sensitive to cyclopamine (ref. 18 and data not shown), we screened several vertebrate cell lines for a sensitive transcriptional response to palmitoyl- and cholesteryl-modified ShhN polypeptide (ShhN_p)²⁰ (Fig. 1a) using a Gli-dependent luciferase reporter²¹. Among several responsive fibroblast cell lines, NIH-3T3 mouse embryonic fibroblasts, which respond with a 20–150-fold induction of luciferase activity (Fig. 1b), were selected for all further studies except those requiring particular genetic backgrounds. Treatment of the cells with cyclopamine completely abolished the response to ShhN_p (Fig. 1c). To confirm the validity of this assay, we analysed the

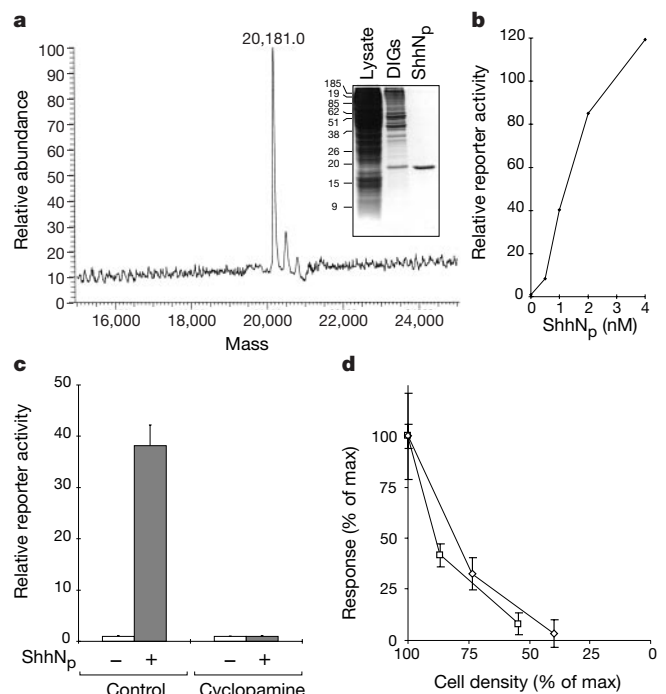


Figure 1 NIH-3T3 fibroblasts respond to ShhN_p and are sensitive to cyclopamine. **a**, Sonic hedgehog signalling domain was purified by isolating detergent-insoluble glycolipid complexes (DIGs) from 293 cells expressing full-length Shh², followed by immunoaffinity chromatography. The purified species corresponds to palmitoyl- and cholesteryl-modified ShhN polypeptide (ShhN_p) from mass spectrometry analysis. Inset, SDS-polyacrylamide gel electrophoresis analysis of lysate, DIGs and purified ShhN_p. The responsiveness of NIH-3T3 cells to Shh (**b**) and cyclopamine (**c**) was analysed by transfection with Gli-luc reporter and Renilla luciferase control followed by treatment with ShhN_p (4 nM or as indicated) and/or cyclopamine (5 μ M) for 2 d. Luciferase activities normalized relative to control. For time response, see Supplementary Information. **d**, The effect of cell density on Shh pathway activity was analysed by plating cultures of stable Shh reporter lines Shh-LIGHT (squares) or SmoA1-LIGHT (diamonds) cells in quadruplicate to 96-well plates in a series of twofold dilutions. The Shh-LIGHT cells were treated with 4 nM ShhN_p, and Gli-luc reporter activity (relative to equally dense Shh-LIGHT control culture) and relative cell density (% of maximum Renilla activity) were assayed after 24 h. Error bars, 1 s.d.

effects of overexpression of known pathway components, treatment with known pathway inhibitors, or both (Table 1). The main findings of *Drosophila* and mouse genetic analyses were confirmed, indicating that NIH-3T3 cells provide a faithful and physiologically meaningful model for analysis of the Shh signalling pathway. Interestingly, for a full response to ShhN_p cells had to be assayed after reaching saturation density (Fig. 1d).

The steroidal nature of cyclopamine and its ability to disrupt cholesterol synthesis or transport^{3,22,23} indicated that it might affect the action of Ptch, which contains an apparent sterol-sensing domain^{1,22}. Having established the general characteristics of Shh response and cyclopamine inhibition in mouse embryonic fibroblasts, we assayed fibroblasts derived from *Ptch*^{-/-} mouse embryos for cyclopamine sensitivity. Mice lacking functional Ptch show widespread transcriptional activation of targets of Shh signalling, including *Ptch* itself²⁴. As β -galactosidase is expressed under the control of the *Ptch* promoter in these cells, its expression can be used to assay the state of Shh pathway activity. Addition of cyclopamine to *Ptch*^{-/-} cells significantly suppressed β -galactosidase expression (Fig. 2a) and the activity of the Gli-luc reporter (see Supplementary Information), indicating that cyclopamine can inhibit Shh pathway activity in the absence of Ptch function. In contrast, cyclopamine failed to prevent pathway activation induced by Gli2 overexpression (Table 1).

These results suggest that Hh-pathway-related tumours associated with loss of Ptch function might respond to treatment with cyclopamine, and that the target of cyclopamine action is likely to be a pathway component that functions between Ptch and the Gli proteins. It is unlikely that Ptch2 is the target of cyclopamine action in *Ptch*^{-/-} cells, as Ptch is the main regulator of Shh pathway activation in embryos²⁴ and Ptch2 activity appears not to be expressed in *Ptch*^{-/-} cells (data not shown).

To investigate further the mechanism of cyclopamine action, we transfected NIH-3T3 cells transiently with both luciferase reporter and Smo complementary DNA, and found that overexpression of Smo in the absence of Shh induces reporter expression about tenfold. This Shh-independent activation of the response pathway can be suppressed by 5 μ M cyclopamine (Fig. 2b), consistent with a target of cyclopamine action downstream of Ptch and with a mechanism that does not involve direct interference with Shh binding. Cyclopamine at this concentration had little effect on

reporter expression induced by the oncogenic Smo mutants W539L¹³ (SmoA1; Fig. 2b) and S537N¹¹ (SmoA2; data not shown); cells from *Ptch* mutant embryos gave similar results (see Supplementary Information).

These results indicate that cyclopamine may act upon Smo, and that activating mutations render Smo proteins resistant. An alternative interpretation would be that activated Smo proteins produce a high abundance of a downstream component and that a high cyclopamine level is required to suppress the increased concentration of this component. This alternative model, however, would predict that intermediate or low levels of pathway activation by oncogenic Smo proteins expressed at low levels should be subject to cyclopamine inhibition; on the contrary, we observe that cyclopamine resistance is sustained under these conditions (Fig. 2b). In addition, cyclopamine resistance is not observed in cells expressing high levels of wild-type Smo activated by maximal Shh stimulation (not shown), again suggesting that cyclopamine does not act upon a component downstream of Smo.

The oncogenic SmoA1 protein has been reported to resist suppression by Ptch²⁵, indicating that oncogenic Smo proteins may not be subject to normal regulation. We found, however, that this resistance is partial. The activating effects of SmoA1 or SmoA2 can be completely inhibited by transfection of a 9-to-1 ratio of a Ptch construct or of *Ptch*-CTD, which encodes a carboxy-terminally deleted protein expressed at a higher level²⁶ (Fig. 2c and data not shown). We also found that SmoA1, thus inhibited by Ptch-CTD, responds well to stimulation by ShhN_p. Under these circumstances,

Table 1 Luciferase activity from a Gli-dependent reporter as induced by combinations of Shh pathway inducing and suppressing treatments

Inducer	Suppressor					
	None	Ptc	Cyclopamine	Forskolin	PKA ^a	Gli3-N
None	—	—	—	—	—	—
ShhN _p	+++	+	—	—	—	—
Smo	+	—	—	—	—	ND
Activated Smo	+++	++	++	—	—	—
Gli2	++	++	++	++	+	—

The indicated constructs were transfected alone or in combination (1:1 ratio) to NIH-3T3 cells, followed by treatment with ShhN_p (4 nM), cyclopamine (5 μ M) or forskolin (100 μ M) for 2 days. Gli3-N, Gli3 truncated at residue 700, generating a repressor form. PKA^a, constitutively active protein kinase A catalytic subunit.

^a Higher dose can completely suppress (see Figs 2, 3).

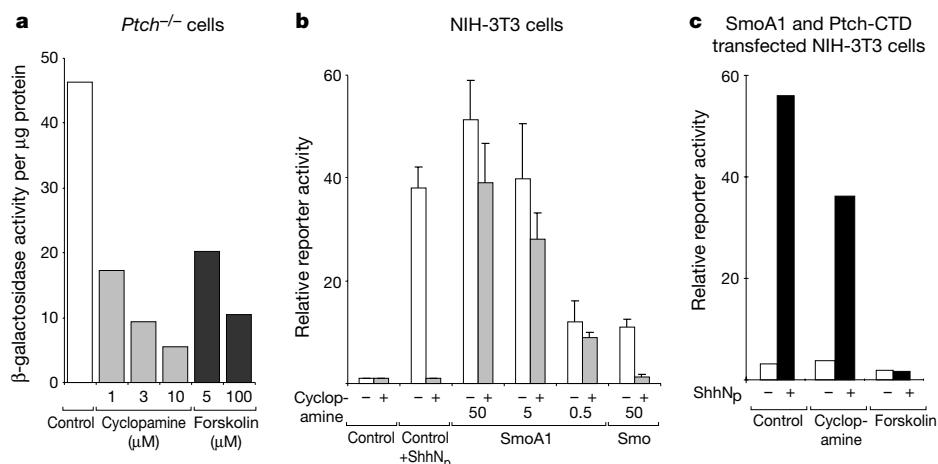


Figure 2 Cyclopamine acts downstream of Ptch by inhibiting Smo activity. **a**, Cyclopamine and forskolin can inhibit Hh pathway activity in *Ptch*^{-/-} cells (measured as expression of β -gal activity from the *Ptch* locus after 3 d of treatment). **b**, Sensitivity of wild-type and oncogenic Smo to cyclopamine was determined by transfecting NIH-3T3 cells with Gli-luc reporter and the w/w ratio of the expression vector indicated, followed by treatment with 5 μ M cyclopamine for 2 d. Error bars, 1 s.d. The leftmost four bars are as in Fig. 1c.

c, Activation of pathway by oncogenic Smo is inhibited by high amounts of Ptch (compare with **b**); ShhN_p restores pathway activity even in the presence of cyclopamine. NIH-3T3 cells were transfected with reporter, Ptch-CTD and SmoA1 (Ptch:Smo DNA ratio 9), and treated with ShhN_p (2 nM), cyclopamine (5 μ M) and/or forskolin (100 μ M) as indicated for 2 d. Data from a representative experiment are shown.

induction of the Gli-responsive reporter is resistant to 5 μ M cyclopamine (Fig. 2c), which would normally abolish Shh signaling. These results indicate that activated Smo molecules in the presence of sufficient Ptch can contribute to an essentially normal, albeit cyclopamine-resistant, response to the Shh signal.

The finding that oncogenic Smo is regulated by high levels of Ptch indicates that it might also be subject to regulation by high levels of cyclopamine. To circumvent the cytotoxic effects of cyclopamine concentrations greater than 10 μ M, we tested several chemically synthesized cyclopamine derivatives. As seen in Fig. 3a, the cyclopamine derivative 3-keto, *N*-aminoethyl aminocaproyl dihydrocin-

namoyl cyclopamine (KAAD-cyclopamine; see Methods), had 10–20-fold higher potency than cyclopamine in inhibition of β -galactosidase expression in $p2^{Ptch-/-}$ cells, with similar or lower toxicity (Fig. 3a). This compound also had greater potency in suppression of ShhN_p-induced pathway activity (Fig. 3b). Importantly, it completely suppressed SmoA1-induced reporter activity at a concentration around tenfold higher than that required for suppression of pathway activation induced by ShhN_p (Fig. 3b).

As activation of the Hh response pathway is associated with

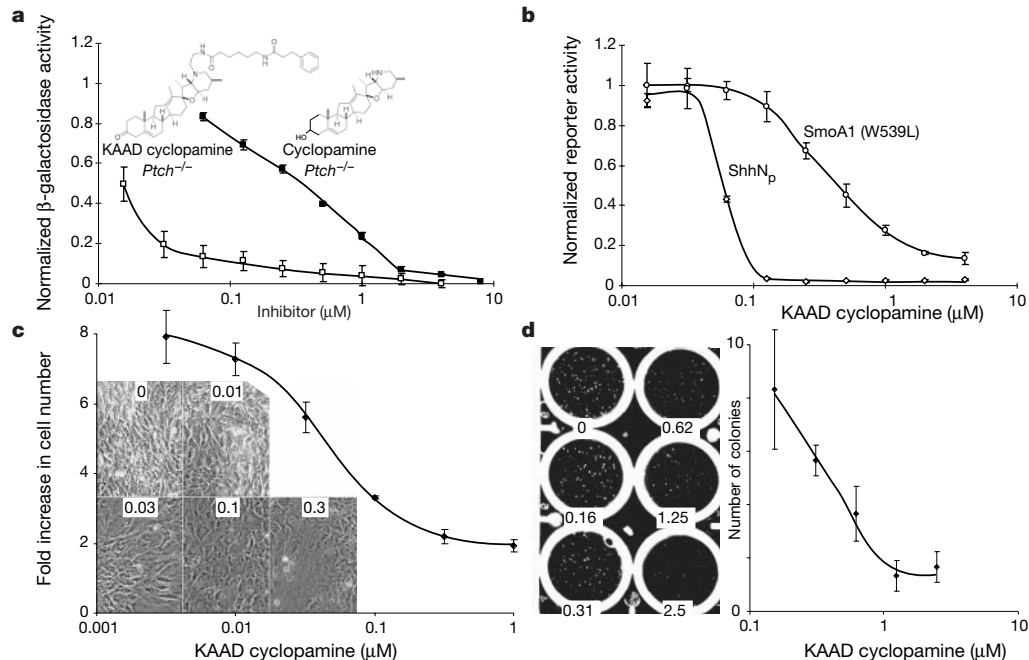


Figure 3 Reversal of pathway activation and abnormal cell growth caused by oncogenic mutations in Ptch and Smo. **a**, Confluent cultures of $p2^{Ptch-/-}$ cells were treated with cyclopamine (filled squares) or KAAD-cyclopamine (open squares) for 2 d. β -galactosidase activity was determined relative to cell mass (treated duplicate plate; Cell Titer 96AQ; Promega). Significant toxicity was not observed. **b**, Shh-LIGHT2 (diamonds) and SmoA1-LIGHT (circles) cells were treated with 4 nM ShhN_p (Shh-LIGHT2) and KAAD cyclopamine (both lines) for 2 d. **c**, $p2^{Ptch-/-}$ cells were cultured in 2% calf serum in the

presence of KAAD cyclopamine for 7 d (triplicate wells) and cell number was determined with a Coulter counter. Inset, micrographs of treated cells. **d**, SmoA1-LIGHT cells were plated in soft agar with the indicated concentrations of KAAD-cyclopamine. Left, ethidium-bromide-stained colonies after 17 d. Right, number of colonies (>150 μ m) per field after 10 d. Error bars, 1 s.d. Relative reporter activities in **a**, **b** are normalized to maximum.

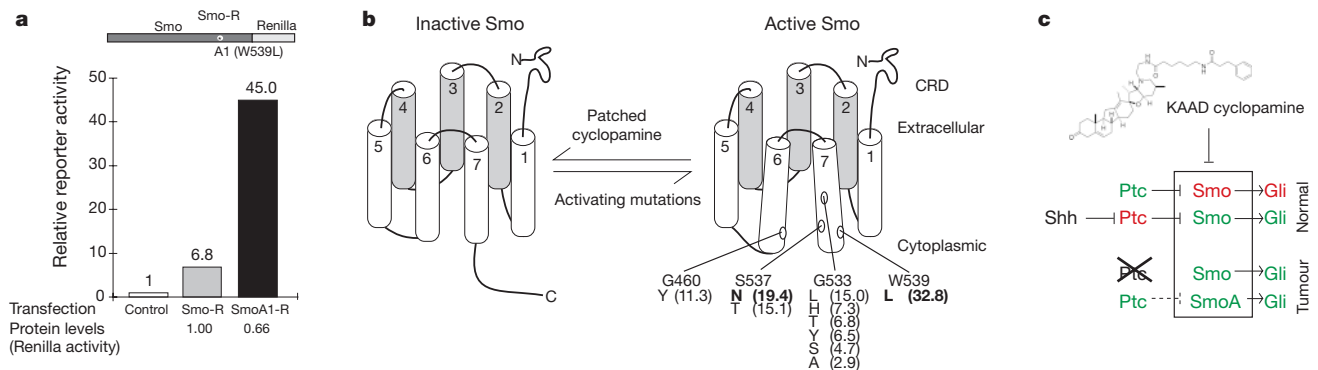


Figure 4 Mechanism of cyclopamine reversal of oncogenic Hh pathway activation. **a**, NIH-3T3 cells were transfected with Gli-luc reporter and a construct encoding the indicated Smo–Renilla luciferase fusion protein. Shh pathway activity and Smo protein levels were measured as firefly and Renilla luciferase activities relative to β -galactosidase transfection control, respectively. Expression levels of epitope-tagged Smo and SmoA1 also were similar (data not shown). **b**, A two-state model of Smo function and cyclopamine action. The states may reflect conformational change (as shown), distinct subcellular localizations or states of covalent modification. Below, summary of oncogenic^{11,13} (bold) and other activating mutations in Smo. The potency of each mutant protein in pathway activation relative to wild-type Smo (value set at 1; ShhN cotransfection, 21.1) is in

parentheses. Relative potencies were determined at 10% w/w effector plasmid. **c**, Cyclopamine inhibition of the Hh response pathway. Under physiological conditions the inactive state of Smo (red) is maintained by active Ptch (green), leading to lack of Gli-mediated transcription (red). Shh ligand activates the pathway by binding to and converting Ptch to its inactive state (red), which in turn permits activity of Smo (green) and results in Gli-mediated transcription (green). In tumours, Gli-mediated transcription is ligand-independent owing to loss of Ptch function (Ptch) or constitutive activation of Smo (SmoA). Cyclopamine inhibits the Hh response pathway and abnormal growth by inhibiting Smo activity.

neoplastic transformation in various types of tumour, we investigated the growth properties of response-activated cells in low serum or soft agar²⁷, conditions generally considered to reveal neoplastic transformation. As seen in Fig. 3c, p2^{Ptch^{-/-}} cell growth in low serum was markedly inhibited by addition of KAAD-cyclopamine, with 50% of maximal inhibition at ~50 nM. We also found that SmoA1-LIGHT cells, a cell line expressing SmoA1 clonally derived from NIH-3T3 cells, can form colonies in soft agar medium. Addition of KAAD-cyclopamine markedly inhibited colony growth, although ~500 nM was required for 50% inhibition (Fig. 3d). This concentration was higher than that required for 50% maximal inhibition of p2^{Ptch^{-/-}} cell growth, consistent with the higher amount of KAAD-cyclopamine required to block activation of the Hh response pathway by oncogenic Smo.

The activity of other signalling pathways headed by seven-transmembrane-domain receptors is determined by the balance between active and inactive forms of the receptor^{28,29}. Consistent with such a model for Smo, the level of Renilla-luciferase- or epitope-tagged Smo in transfected cells is not significantly affected by the presence of an oncogenic mutation, despite greatly elevated reporter activity (Fig. 4a). These results indicate that oncogenic Smo may have a higher intrinsic ability to activate the pathway, and that wild-type Smo may exist in a balance between active and inactive forms^{28,29}. Whereas cyclopamine and Ptch activities would shift this balance towards the inactive state, oncogenic mutations might exert the opposite effect, and this could account for the higher levels of Ptch and cyclopamine activity required to suppress oncogenic Smo. To test this model, we identified eight additional mutations in an extensive mutagenesis screen for activated Smo proteins (Fig. 4b). All the activating Smo mutants were cyclopamine resistant, and the level of resistance positively correlated with increased potency (see Supplementary Information), consistent with a model in which cyclopamine affects the balance between activity states of Smo (Fig. 4b).

The transition of G-protein-coupled receptors with seven transmembrane (TM) domains from the inactive to the active state is thought to involve a conformational shift in which the cytoplasmic ends of the TM6 and TM7 helices tilt outwards, exposing a binding pocket for a downstream signalling molecule, the G α subunit²⁹. Our mutational analysis of Smo is consistent with such a conformational shift, as eight of the Smo-activating mutations introduce bulkier side chains at G533 and S537. Assuming an α -helical conformation for TM7, these substitutions all protrude from the same face of the helix. Furthermore, as multiple distinct substitutions at each of these two residues result in activation (Fig. 4b), Smo would appear to be activated by alteration of helix-packing interactions rather than by creation or disruption of a single critical interaction. A conformational shift of the TM7 helix with respect to other TM helices is suggestive of the type of conformational transition postulated for activation of G-protein-coupled receptors and raises the possibility that conformation-based transduction has been conserved in the evolution of seven-transmembrane-domain receptors.

Activation of the Hh response pathway has been linked to several types of human tumour. For example, patients with basal cell nevus syndrome (also termed Gorlin syndrome), an autosomal dominant disorder associated with heterozygous loss-of-function mutations in *PTCH*, display increased incidence of many tumours, most notably basal cell carcinoma (BCC), medulloblastoma, rhabdomyosarcoma and fibrosarcoma⁶⁻⁸. In addition, loss-of-function mutations in *PTCH* or activating mutations in *SMO* are found in around 40% of sporadic BCC and 25% of primitive neuroectodermal tumours^{6,9-13}. Our results indicate that cyclopamine inhibits the Shh pathway by antagonizing Smo and that the activation of the Hh response pathway by either type of oncogenic mutation is blocked by cyclopamine or its derivatives (Fig. 4c). Levels of cyclopamine or the related compound jervine required to phenocopy the embryonic malformations in *Shh^{-/-}* embryos are tolerated by pregnant females

of various species, indicating that it might be possible to use these compounds or their derivatives without severe toxicity in non-pregnant adults to reverse activation of the Shh response pathway for therapeutic purposes. □

Methods

Synthesis of KAAD-cyclopamine and purification of ShhN_p

KAAD-cyclopamine was synthesized from cyclopamine by a route to be described elsewhere (J.K.C. and P.A.B., manuscript in preparation). Cholesterol- and palmitate-modified mouse Sonic hedgehog signalling domain ShhN_p was purified as described in Supplementary Information (see also Fig. 1a).

cDNA cloning, mutagenesis and constructs

A mouse full-length *Smo* cDNA was isolated using standard methods on the basis of homology to rat and human *Smo* sequences. Of two described activating mutations in human *Smo*^{13,25}, one (SmoM2; corresponds to W539L in mSmo and referred to here as SmoA1) activated mouse Smo and the other (SmoM1; corresponding to R565Q) did not. For mutagenesis screens, Ala substitutions were introduced at more than 20 conserved or hydrophilic residues within Smo transmembrane segments; in addition, selected residues were changed to 16 different amino acids (all except W, M, K and E) using a degenerate oligonucleotide and screened by transfection of a pool of clones. We secondarily screened 53 (G533) or 37 (G460) individual clones to identify mutations at these positions that activate Smo. Expression constructs used: pRK5 for full-length mouse Ptch, C-terminally truncated Ptch-CTD, Gli3(1–700) repressor¹⁹ and constitutively active PKA¹⁸ cDNAs, and pGE (transient transfections) or pcDNA3.1+hygro (stable lines; Invitrogen) for the various *Smo* cDNAs. pGE was derived from pEGFP-C1 (Clontech) by removal of the enhanced green fluorescent protein cDNA. Renilla luciferase was fused to the C terminus of Smo. These fusion protein constructs had similar activity to corresponding untagged constructs in the NIH-3T3 assay.

Cell lines

Primary fibroblasts lacking functional Ptch (*Ptch^{-/-}*) and a clonal subline (p2^{Ptch^{-/-}}) were derived from *Ptch^{-/-}* mouse embryos. The NIH-3T3 cell clones Shh-LIGHT and Shh-LIGHT2 stably incorporating the Gli-luc reporter and TK-Renilla vectors were established by co-transfection with vector encoding G418 resistance (pSV-Neo). SmoA1-LIGHT cells are a clonal subline of Shh-LIGHT expressing oncogenic Smo (verified by immunoblotting).

Transfection and reporter assays

Confluent cultures of NIH-3T3 cells were plated at 1:6 dilution to 24- or 96-well plates. On the next day, we transfected the cells with Renilla luciferase (pRL-TK or pRL-SV40; Clontech) or β -galactosidase transfection control (10% w/w DNA), Gli-luc reporter (40%) and the constructs indicated (50%) using Fugene 6 (Roche) transfection reagent (250 ng (24-well plate) or 100 ng (96-well plate) DNA per well, 3:1 ratio (v/w) of reagent to DNA). After the cells had reached saturation density (1–2 days), they were changed to low-serum medium (0.5% bovine calf serum) and treated with the reagents indicated for 1–2 days. All reporter assays were normalized for transfection efficiency using Renilla luciferase or β -galactosidase transfection control (luminometric detection, Promega dual luc; Tropix galacto-light).

Received 12 May; accepted 6 July 2000.

- Goodrich, L. V. & Scott, M. P. Hedgehog and patched in neural development and disease. *Neuron* **21**, 1243–1257 (1998).
- Ingham, P. W. Transducing hedgehog: the story so far. *EMBO J.* **17**, 3505–3511 (1998).
- Cooper, M. K., Porter, J. A., Young, K. E. & Beachy, P. A. Plant-derived and synthetic teratogens inhibit the ability of target tissues to respond to Sonic hedgehog signaling. *Science* **280**, 1603–1607 (1998).
- Incardona, J. P., Gaffield, W., Kapur, R. P. & Roelink, H. The teratogenic *Veratrum* alkaloid cyclopamine inhibits sonic hedgehog signal transduction. *Development* **125**, 3553–3562 (1998).
- Chiang, C. *et al.* Cyclopamine and defective axial patterning in mice lacking *Sonic hedgehog* gene function. *Nature* **383**, 407–413 (1996).
- Hahn, H., Wojnowski, L., Miller, G. & Zimmer, A. The patched signaling pathway in tumorigenesis and development: lessons from animal models. *J. Mol. Med.* **77**, 459–468 (1999).
- Hahn, H. *et al.* Mutations of the human homolog of *Drosophila* patched in the nevoid basal cell carcinoma syndrome. *Cell* **85**, 841–851 (1996).
- Johnson, R. L. *et al.* Human homolog of *patched*, a candidate gene for the basal cell nevus syndrome. *Science* **272**, 1668–1671 (1996).
- Gailani, M. *et al.* The role of the human homologue of *Drosophila* patched in sporadic basal cell carcinomas. *Nature Genet.* **14**, 78–81 (1996).
- Pietsch, T. *et al.* Medulloblastomas of the desmoplastic variant carry mutations of the human homologue of *Drosophila* patched. *Cancer Res.* **57**, 2085–2088 (1997).
- Reifenberger, J. *et al.* Missense mutations in SMOH in sporadic basal cell carcinomas of the skin and primitive neuroectodermal tumors of the central nervous system. *Cancer Res.* **58**, 1798–1803 (1998).
- Raffel, C. *et al.* Sporadic medulloblastomas contain *PTCH* mutations. *Cancer Res.* **57**, 842–845 (1997).
- Xie, J. *et al.* Activating *Smoothed* mutations in sporadic basal-cell carcinoma. *Nature* **391**, 90–92 (1998).
- Gibbs, J. B. Mechanism-based target identification and drug discovery in cancer research. *Science* **287**, 1969–1973 (2000).
- Keeler, R. F. & Binns, W. Teratogenic compounds of *Veratrum californicum* (Durand). V. Comparison of cyclopamine effects of steroidal alkaloids from the plant and structurally related compounds from other sources. *Teratology* **1**, 5–10 (1968).

16. Aza-Blanc, P., Ramirez-Weber, F.-A., Laget, M.-P., Schwartz, C. & Kornberg, T. B. Proteolysis that is inhibited by Hedgehog targets cubitus interruptus protein to the nucleus and converts it to a repressor. *Cell* **89**, 1043–1053 (1997).
17. Methot, N. & Basler, K. Hedgehog controls limb development by regulating the activities of distinct transcriptional activator and repressor forms of Cubitus interruptus. *Cell* **96**, 819–831 (1999).
18. Chen, C. H. et al. Nuclear trafficking of Cubitus interruptus in the transcriptional regulation of Hedgehog target gene expression. *Cell* **98**, 305–316 (1999).
19. Wang, B., Fallon, J. & Beachy, P. Hedgehog-regulated processing of Gli3 produces an anterior/posterior repressor gradient in the developing vertebrate limb. *Cell* **100**, 423–434 (2000).
20. Pepinsky, R. B. et al. Identification of a palmitic acid-modified form of human Sonic hedgehog. *J. Biol. Chem.* **273**, 14037–14045 (1998).
21. Sasaki, H., Hui, C.-C., Nakafuku, M. & Kondoh, H. A binding site for Gli proteins is essential for HNF3 β floor plate enhancer activity in transgenics and can respond to Shh *in vitro*. *Development* **124**, 1313–1322 (1997).
22. Beachy, P. A. et al. Multiple roles of cholesterol in hedgehog protein biogenesis and signaling. *Cold Spring Harb. Symp. Quant. Biol.* **62**, 191–204 (1997).
23. Lange, Y., Ye, J., Rigney, M. & Steck, T. L. Regulation of endoplasmic reticulum cholesterol by plasma membrane cholesterol. *J. Lipid. Res.* **40**, 2264–2270 (1999).
24. Goodrich, L. V., Milenkovic, L., Higgins, K. M. & Scott, M. P. Altered neural cell fates and medulloblastoma in mouse patched mutants. *Science* **277**, 1109–1113 (1997).
25. Murone, M., Rosenthal, A. & de Sauvage, F. J. Sonic hedgehog signaling by the patched-smoothed receptor complex. *Curr. Biol.* **9**, 76–84 (1999).
26. Fuse, N. et al. Sonic hedgehog protein signals not as a hydrolytic enzyme but as an apparent ligand for patched. *Proc. Natl Acad. Sci. USA* **96**, 10992–10999 (1999).
27. Shin, S. I., Freedman, V. H., Risser, R. & Pollack, R. Tumorigenicity of virus-transformed cells in nude mice is correlated specifically with anchorage independent growth *in vitro*. *Proc. Natl Acad. Sci. USA* **72**, 4435–4439 (1975).
28. Bond, R. A. et al. Physiological effects of inverse agonists in transgenic mice with myocardial overexpression of the beta 2-adrenoceptor. *Nature* **374**, 272–276 (1995).
29. Bourne, H. R. How receptors talk to trimeric G proteins. *Curr. Opin. Cell Biol.* **9**, 134–142 (1997).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank P. A. Cole, D. Leahy, J. Nathans, R. Reed and G. Seydoux for critical review of the manuscript, W. Gaffield for *Veratrum* extracts and K. Young for technical assistance. J.T. was supported by the Finnish Academy and the European Molecular Biology Organization. J.K.C. is a recipient of a postdoctoral fellowship from the Damon Runyon-Walter Winchell Cancer Research Fund and M.K.C. was supported by a postdoctoral fellowship from the Howard Hughes Medical Institute and a career development award from the Burroughs Wellcome Fund. P.A.B. and M.P.S. are investigators of the Howard Hughes Medical Institute.

Correspondence and requests for materials should be addressed to P.A.B. (e-mail: pbeachy@jhmi.edu).

The APC tumour suppressor has a nuclear export function

Rina Rosin-Arbesfeld, Fiona Townsley & Mariann Bienz

Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

The adenomatous polyposis coli (APC) protein is mutated in most colorectal tumours¹. Nearly all APC mutations are truncations, and many of these terminate in the mutation cluster region located halfway through the protein^{2–4}. In cancer cells expressing mutant APC, β -catenin is stabilized^{5,6} and translocates into the nucleus to act as a transcriptional co-activator of T-cell factor⁷. During normal development, APC also promotes the destabilization of β -catenin and *Drosophila* Armadillo^{8–11}. It does so by binding to the Axin complex which earmarks β -catenin/Armadillo for degradation by the proteasome pathway¹². APC has a regulatory role in this process^{13,14}, which is poorly understood. Here we show that APC contains highly conserved nuclear export signals 3' adjacent to the mutation cluster region that enable it to exit from the nucleus. This ability is lost in APC mutant cancer cells, and we provide evidence that β -catenin accumulates in the nucleus as a result. Thus, the ability of APC to exit from the nucleus appears to be critical for its tumour suppressor function.

Adenomatous polyposis coli proteins are found in several subcellular compartments of mammalian and *Drosophila* cells including the cytoplasm, nucleus and adhesive cadherin/catenin junctions^{9,10,15}. To identify the targeting domains for these compartments, we tagged various fragments of the ubiquitously expressed *Drosophila* APC, called E-APC/dAPC2, with green fluorescent protein (GFP) and expressed these in transgenic fly embryos and in monkey COS cells. The subcellular distribution of GFP-E-APC is indistinguishable from that of endogenous E-APC in embryos⁹ (data not shown). In COS cells transfected with GFP-E-APC, we see green fluorescence in the cytoplasm, concentrated at the plasma membrane (to be described elsewhere), but also some in the nucleus (Fig. 1a). Unexpectedly, an amino-terminal fragment of E-APC (ARDcore; Fig. 2a) accumulates in the nucleus (Fig. 1b). Evidently, E-APC is capable of entering the nucleus by means of its N terminus. We have not studied this further, but we note that this N terminus spans the highly conserved Armadillo repeat domain (ARD). β -catenin contains a closely related ARD that mediates its nuclear import independently of the Ran/Importin machinery^{16,17}.

In contrast, carboxy-terminal fragments of E-APC (Cterm1 and 2; Fig. 2a) are efficiently excluded from the nucleus (Fig. 1c), more so than the full-length protein. Thus, the C terminus of E-APC either contains a cytoplasmic anchoring domain or an efficient nuclear export signal (NES). To distinguish between these possibilities, we tested our GFP constructs by treating transfected cells with leptomycin B (LMB), a highly specific drug that inhibits nuclear export by directly blocking the nuclear export receptor CRM1 (refs 18–20). Indeed, this results in even distribution of Cterm1 and Cterm2 throughout cytoplasm and nucleus (Fig. 1d).

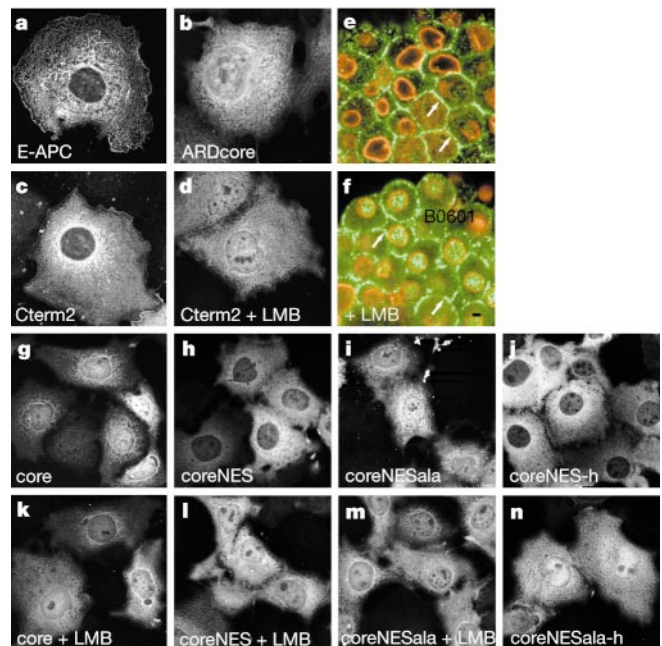


Figure 1 Nuclear export sequences within *Drosophila* and human APC. **a–d**, COS cells transfected with GFP-tagged constructs as indicated, without (**a–c**) or with (**d**) LMB. Cterm1 was excluded from nuclei at least as efficiently as Cterm2 (**c**); this exclusion, like that of Cterm2 (**d**), was blocked by LMB. **e, f**, Face-on views of 4–5-hour-old *Drosophila* embryos, stained with anti-E-APC (green) outlining the apical cellular junctions (arrows), and anti-lamin (red) outlining the nuclear envelopes. Note the nuclear accumulation of E-APC after LMB treatment (**f**). **g–n**, COS cells transfected with coreNES constructs. **g, k**, ARDcore; **h, l**, coreNES from E-APC 20R4; **i, m**, coreNESala from E-APC 20R4; **j, n**, coreNES from human APC 20R3; **o, p**, coreNESala from human APC 20R3. All coreNES constructs described in the text behaved essentially the same. Cells in **k–m** were treated with LMB; the staining patterns of the human constructs (**j, n**) after LMB treatment are similar to those in **k–m**.

Full-length E-APC also accumulates to some extent in the nucleus after LMB treatment (not shown). Notably, endogenous E-APC is retained efficiently in nuclei of LMB-treated *Drosophila* embryos (Fig. 1f, compare with 1e). These results indicate the presence of an NES in the C terminus of E-APC.

To identify this, we scanned through the C-terminal sequence of E-APC for matches to the leucine-rich NES consensus sequence (LxxLxΦ, Φ being L, I, M or V). We found two matches in intriguing positions, namely in the so-called '20-amino-acid repeat' 3 and 4 (20R3, 20R4) (Fig. 2a, red arrows). The 20Rs are highly conserved motifs (Fig. 2, black bars) which in human APC bind β-catenin^{5,21,22}. These putative NESs in 20R3 and 20R4 are conserved in all known APC proteins. Human APC contains an additional NES match in 20R7 (Fig. 2b, red arrow). Two further matches were found in both proteins (Fig. 2, black arrows, arrowhead). We tested the R20-linked NESs from E-APC and human APC, by fusing each individually to the ARDcore (coreNES). We also generated mutant versions which bear alanine substitutions of the conserved NES residues (coreNESsala), as well as a control alanine mutant (coreNEScon) and a minimal NES construct (coreNESmin; see Methods).

We found that all coreNES fusions are efficiently excluded from nuclei of transfected COS cells (Fig. 1h, j). Similarly, cells transfected with coreNESmin and coreNEScon show barely any

green fluorescence in the nuclei (data not shown). In contrast, green fluorescence from coreNESsala mutants is evenly spread throughout cytoplasm and nuclei of transfected cells (Fig. 1i, n). These mutants are thus indistinguishable from ARDcore (Fig. 1g). LMB treatment abolishes the nuclear exclusion of all coreNES fusions (Fig. 1l), but neither affects the subcellular distribution of ARDcore (Fig. 1k) nor that of any coreNESsala mutant (Fig. 1m). We conclude that the 20R-linked NESs from human and *Drosophila* APC function as nuclear export signals.

Intriguingly, the 20R3 NES overlaps the mutation cluster region (MCR) of human APC. Plotting the codon positions of 315 somatic mutations from colorectal tumours²⁻⁴ revealed a fairly even spread throughout the MCR, with known hot-spots, up to an abrupt 3' border immediately upstream of the 20R3 NES (codon 1,506; Fig. 2b). Only 5 mutations fall within the 72 codons between this border and the furthest 5' Axin-binding motif (codon 1,570; Fig. 2b) which is thought to be critical for the tumour suppressor function of APC^{23,24}. Out of 573 somatic colorectal tumour mutations compiled in the APC database, only 16 (2.8%) fall 3' to this border. Germline mutations 3' to codon 1,465 usually lead to fewer polyps than those located more upstream, and those 3' to 1,578 are associated with attenuated polyposis²⁵. Note that these APC mutations result in protein truncations. Therefore, this observed 3' border indicates a

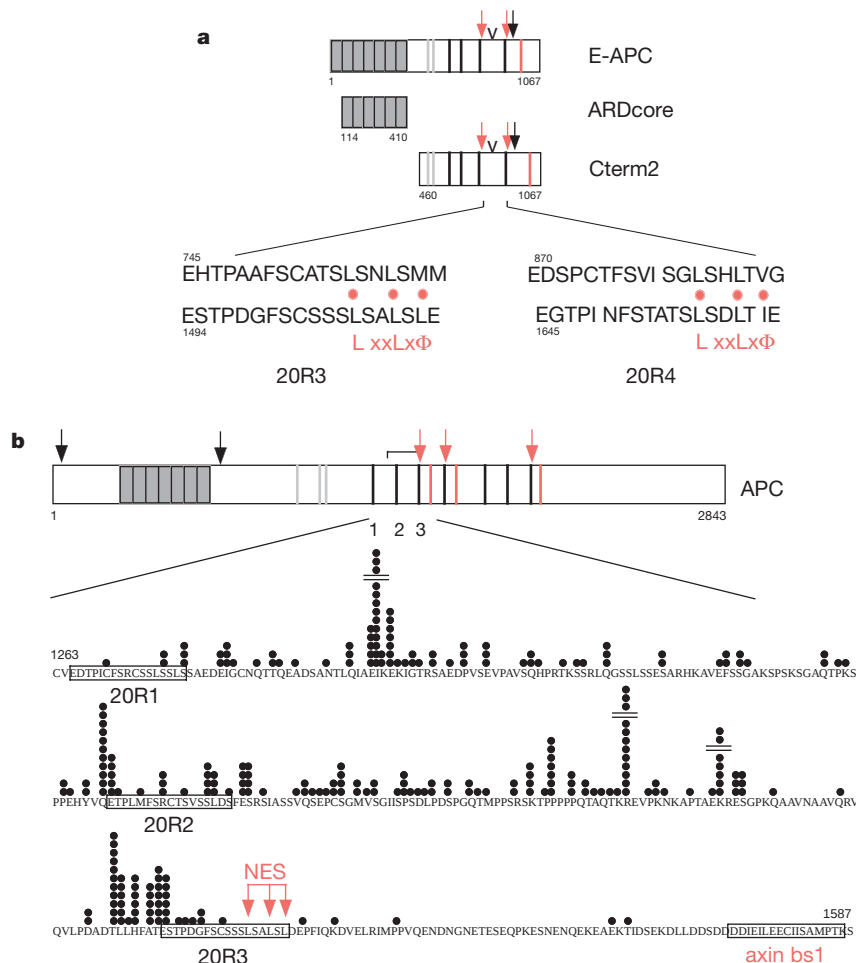


Figure 2 Maps of APC proteins and positions of NESs relative to the MCR. **a**, Top, E-APC with conserved domains (black bars, 20Rs; grey bars, 15Rs; red bar, Axin-binding site; grey, ARD), and maps of ARDcore and Cterm2 (Cterm1 is the same as Cterm2, but terminates at codon 908, 5' to the Axin-binding site). Red arrows mark functional NESs. An untested NES candidate is indicated by black arrow, a non-functional NES candidate by arrowhead. Bottom, sequences of 20R3 and 20R4 of E-APC (above) and human APC

(underneath), with conserved NES residues marked by red dots. **b**, Human APC with conserved domains, and functional and untested NESs marked as in **a**. The MCR is bracketed and expanded below to show the codon positions of 315 somatic truncation mutations from colorectal tumours²⁻⁴ (dots, double-bars indicate 10 additional mutations at each hot-spot; see also APC Mutation Database (<http://perso.curie.fr/Thierry.Soussi/APC.html>)). Note abrupt 3' border of mutations immediately upstream of the 20R3 NES.

strong selection against the presence of the 20R-linked NESs in cancers, implicating the ability of APC to exit from the nucleus in its tumour suppressor function.

To gain further support for this, we examined the subcellular distribution of endogenous APC in HCT116 colon cancer cells which contain wild-type APC, and in SW480 cells whose resident truncated APC lacks all 20R-linked NESs^{6,26}, using an antiserum raised against a central fragment of human APC²⁷. This revealed that HCT116 cells contain mostly cytoplasmic APC, some of which is associated with the plasma membrane, and very little is seen in the nucleus (Fig. 3a). In contrast, in many SW480 cells, the truncated APC is predominantly nuclear (Fig. 3d, arrow); in some cells, it is spread evenly throughout nucleus and cytoplasm (Fig. 3d). Evidently, this APC truncation has lost most of its nuclear export function, suggesting that its N-terminal NES matches do not significantly contribute to this function in the mutant protein. Interestingly, these subcellular distributions of APC are largely mirrored by β -catenin (Fig. 3c, f): in HCT116 cells, β -catenin is mostly associated with the plasma membrane, being barely detectable elsewhere (Fig. 3b), whereas in many SW480 cells, β -catenin is concentrated in nuclei (Fig. 3e, arrow). This suggests that the subcellular distribution of β -catenin is a consequence of that of APC.

Minimal 20R-containing APC fragments complement the APC loss in mutant colon cancer cells by reducing their high β -catenin levels^{5,22,24}. We asked whether the nuclear export function of APC is critical for this complementation. We thus generated NES-less APC mutants by introducing all the above-described NESala substitutions into a GFP-tagged central fragment of human APC (HC) to generate HCala, and into full-length E-APC (E-APCala) as E-APC also reduces β -catenin in SW480 cells²⁸. As expected, wild-type HC efficiently exits from nuclei (Fig. 4a, arrow; and 4c), and strongly reduces nuclear β -catenin in transfected SW480 cells (Fig. 4b, arrow). We observed some variability of this effect, for example, cells with low HC levels tended to retain cytoplasmic β -catenin, but we rarely saw a transfected cell whose β -catenin was higher in the nucleus than in the cytoplasm. LMB treatment causes nuclear retention of HC (Fig. 4d, arrows, Fig. 4f), and also attenuates the reduction of nuclear β -catenin by HC in that many HC-transfected cells had more β -catenin in the nuclei than in the cytoplasm (Fig. 4e, arrows). Similarly, HCala, typically retained in nuclei of transfected cells (Fig. 4g, arrow; 4i), is compromised in its ability to reduce their nuclear β -catenin levels (Fig. 4h, arrow). However, the cytoplasmic β -catenin levels on the whole were still reduced in LMB-treated and HCala-transfected cells, indicating that the cytoplasmic APC in these cases retains the ability to destabilize β -catenin (note that

HCala most probably still binds β -catenin²² and Axin^{13,14}). Similar results were obtained with E-APCala which is significantly less active in reducing nuclear β -catenin than its wild-type counterpart (data not shown).

To quantitate the activities of HC and HCala in reducing nuclear β -catenin, we used a transcriptional read-out based on a luciferase reporter linked to T-cell factor (TCF)-binding sites (TOPFLASH; the FOPFLASH control contains mutant TCF-binding sites)⁷. As expected, the high luciferase activity of mock-transfected SW480 cells was much reduced by HC, but less so by HCala (Fig. 4j). A partially mutant HCala that retained the 20R4 and 20R7 NESs (HCala2) was nearly as active as wild-type HC in reducing luciferase activity, whereas a mutant retaining only the 20R7 NES (HCala1) was less active (Fig. 4j). These results confirm the functional importance of the 20R-linked NESs of APC in reducing the nuclear β -catenin in APC mutant cancer cells.

We have shown that APC proteins contain highly conserved and functional nuclear export signals. Three lines of evidence implicate

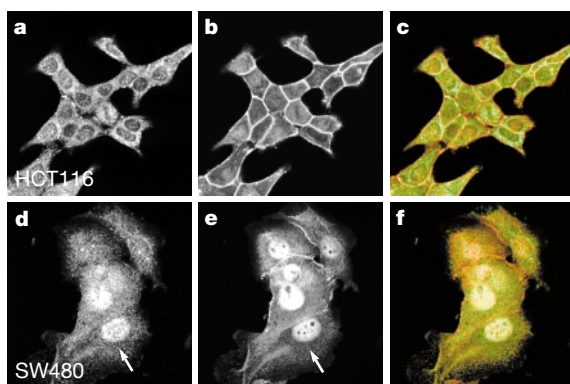


Figure 3 Wild-type and mutant APC in colon cancer cells. HCT116 (a–c) or SW480 (d–f) cancer cells, stained with anti-APC (a, d; green in c, f) and anti- β -catenin (b, e; red in c, f). Wild-type APC is mostly cytoplasmic, and β -catenin membrane-associated in HCT116 cells. In many SW480 cells, the mutant APC is predominantly nuclear (d, arrow), which is mirrored by β -catenin (e, arrow).

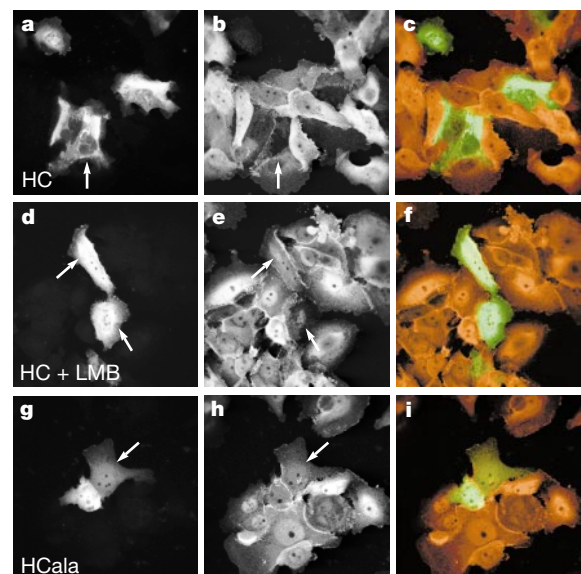


Figure 4 Complementation tests in APC mutant cancer cells. a–i, SW480 cells transfected with GFP-tagged HC (a–f) or HCala (g–i), stained with anti- β -catenin (b, e, h; red in c, f, i); cells in (d–f) were treated with LMB. Arrow in b points to a cell with reduction of nuclear and cytoplasmic β -catenin, arrows in e and h point to cells with nuclear β -catenin. j, Transcriptional read-outs of nuclear β -catenin in SW480 cells transfected with GFP (Mock), HC, HCala, HCala1 or HCala2 (with one or two 20R NESs retained, respectively; see text). Grey columns, TOPFLASH; black columns, FOPFLASH. HCala is significantly less active than HC (or HCala2) in reducing transcriptional activity of nuclear β -catenin.

the ability of APC to exit from the nucleus in its tumour suppressor function: the sharp 3' border of APC truncation mutations, the nuclear accumulation of truncated APC (lacking 20R-linked NESs) in APC mutant cancer cells, and the compromised ability of NES-less APC to reduce nuclear β -catenin in these cells. The nuclear export function of APC appears to be the 5'-most tumour suppressor function within the protein. The ability of APC to bind Axin to destabilize β -catenin, which is also clearly critical for its tumour suppressor function^{23,24}, is encoded slightly downstream, and additional functions may reside in its C terminus.

We have proposed that APC may shuttle β -catenin/Armadillo from the nucleus and cytoplasm to the junctional compartment where the Axin complex appears to be anchored²⁹. Our work provides evidence for this putative shuttling function of APC as the subcellular distribution of β -catenin mirrors that of APC in wild-type and APC mutant cancer cells. However, the nuclear β -catenin in the cancer cells may not simply be a consequence of the loss of APC-mediated export. Perhaps β -catenin is positively trapped in the nuclei by the mutant APC. Nuclear trapping of β -catenin by truncated APC may provide a mechanistic explanation for the striking mutation pattern observed in colorectal tumours, which reveals a strong selection for 20R1 to be retained in at least one of the two mutant APC alleles⁴. □

Methods

Constructs and luciferase assays

All constructs were inserted into pEGFP-C2 (Clontech), producing N-terminally tagged GFP fusions. Each construct (except HC and HCal, see below) also contains a triple haemagglutinin (HA) tag inserted between GFP and APC coding sequence. In coreNES, the following NES-encoding sequences were inserted between the HA tag and the ARDcore (the first of each pair is human, the second *Drosophila*; bold NES residues were substituted to alanine in coreNESal; in coreNEScon, underlined conserved non-NES residues were substituted by alanine): 20R3, ESTPDGFCSSSLALSLEDEP, EHTPAAFSCATSLSNLSM MDD; 20R4, EGTPINFSTATSLSDLTIESP, EDSPCTFSVIGLSHLTVGSA; 20R7, EDTPVCFSRNSLSLSLSDSE. The coreNESmin from *Drosophila* 20R4 contains ISGLSH LTVGSA (bold residues substituted by alanine in the mutant version). We also fused the putative NES from *Drosophila* pseudoR20 (Fig. 2a, arrowhead) to ARDcore, but this NES candidate (EDTTAVLSKAPSNSDSLILSPND) was non-functional.

For the complementation assays in SW480 cells, a central fragment from human APC (codons 1,379–2,080) was N-terminally tagged with GFP (HC; see above). The above described alanine substitutions of 20R-linked NESs were introduced into HC, and into GFP-E-APC, using the Quik-Change Site-Directed Mutagenesis Kit (Stratagene). HCal1 and HCal2 are partial mutants retaining one and two NESs, respectively. All constructs were verified by sequencing.

pTOPFLASH and pFOPFLASH were used for transcriptional read-out assays of nuclear β -catenin in SW480 cells⁷. pRL-CMV served as an internal control, and luciferase assays were performed with the Dual Luciferase Reporter Assay System (Promega). Relative luciferase activities ($\times 100$) were obtained by dividing TOP- or FOPFLASH values by pRL-CMV values (Fig. 4j). TOPFLASH values reflect averages of 2–4 independent transfections; their standard deviations are given.

Tissue culture, fly embryos and immunofluorescence

Monkey COS cells were grown in DMEM medium (supplemented with 10% fetal calf serum), and transfected with FuGENE 6 Transfection Reagent (Roche). SW480 and HCT116 cells were grown in Leibovitz's L15 medium (with 10% fetal calf serum), and transfected with Lipofectamine (Life Technologies). DNA at 0.4 or 0.8 μ g was used per well for transfection of COS or SW480 cells (in 35-mm culture wells), respectively. Transfected cells were collected for analysis after 24–48 h. For the stainings in Figs 1 and 3, cells were fixed with chilled methanol for 10 min at -20°C . For the stainings in Fig. 4, cells were fixed with 4% paraformaldehyde (freshly prepared, in phosphate buffered saline) for 30 min, permeabilized with 0.1% Triton X-100 and blocked with bovine serum albumin for 15 min each, and subsequently incubated at room temperature with primary and secondary antibody for 60 and 30 min, respectively. COS cells were treated with LMB (50 ng ml⁻¹); provided by M. Yoshida) for 1–2 h, 24 h after transfection, SW480 cells for 2 h, 48 h after transfection.

Drosophila embryos were fixed and stained as described⁹. For drug treatments, 0–6-hour-old embryos were permeabilized in octane³⁰ and subsequently incubated for 60 min with 80 ng ml⁻¹ LMB. Controls were octane-permeabilized and incubated in solvent alone (<1% ethanol).

We used the following antibodies: rabbit anti-E-APC (1:10,000)⁹, rabbit anti-APC (provided by I. Näthke; 1:700)²⁷, mouse anti-lamin (gift from M. Frasch, provided by A. Gonzales-Reyes; 1:150), mouse anti- β -catenin (1:500; Transduction Laboratories), Alexa Red and Green (1:500; Molecular Probes). Images were collected on an MRC 1024 confocal microscope.

Some heterogeneity was observed with the NES constructs, and with the SW480 complementation assays. Qualitative analysis was thus backed up by quantitative evaluation of 10–20 randomly chosen fields in which each individual healthy cell was scored. Ratios of nuclear to cytoplasmic fluorescence levels were determined.

Expression levels of GFP fusions were checked by western blotting, essentially as described²⁴ (see Supplementary Information). The following antibody dilutions were used: anti-APC (see above), 1:2,000; rat anti-HA (clone 3F10, Roche), 1:700; anti- α -tubulin (Sigma T9026), 1:500.

Received 1 June; accepted 18 July 2000.

- Kinzler, K. W. & Vogelstein, B. Lessons from hereditary colorectal cancer. *Cell* **87**, 159–170 (1996).
- Nagase, H. & Nakamura, Y. Mutations of the APC (Adenomatous polyposis coli) gene. *Hum. mutat.* **2**, 425–434 (1993).
- Miyaki, M. *et al.* Characteristics of somatic mutation of the Adenomatous polyposis coli gene in colorectal tumors. *Cancer Res.* **54**, 3011–3020 (1994).
- Lamlum, H. *et al.* The type of somatic mutation at APC in familial adenomatous polyposis is determined by the site of the germline mutation: a new facet to Knudson's 'two-hit' hypothesis. *Nature Med.* **5**, 1071–1075 (1999).
- Munemitsu, S., Albert, I., Souza, B., Rubinfeld, B. & Polakis, P. Regulation of intracellular β -catenin levels by the adenomatous polyposis coli (APC) tumor-suppressor protein. *Proc. Natl Acad. Sci. USA* **92**, 3046–3050 (1995).
- Morin, P. J. *et al.* Activation of β -catenin-TCF signaling in colon cancer by mutations in β -catenin or APC. *Science* **275**, 1787–1790 (1997).
- Korinek, V. *et al.* Constitutive transcriptional activation by a β -catenin-Tcf complex in APC^{-/-} colon carcinoma. *Science* **275**, 1784–1787 (1997).
- Ahmed, Y., Hayashi, S., Levine, A. & Wieschaus, E. Regulation of armadillo by a *Drosophila* APC inhibits neuronal apoptosis during retinal development. *Cell* **93**, 1171–1182 (1998).
- Yu, X., Waltzer, L. & Bienz, M. A new *Drosophila* APC homologue associated with adhesive zones of epithelial cells. *Nature Cell Biol.* **3**, 144–151 (1999).
- McCartney, B. M. *et al.* *Drosophila* APC2 is a cytoskeletonally-associated protein that regulates wingless signaling in the embryonic epidermis. *J. Cell Biol.* **146**, 1303–1318 (1999).
- Farr, G. H. *et al.* Interaction among GSK-3, GBP, axin, and APC in *Xenopus* axis specification. *J. Cell Biol.* **148**, 691–702 (2000).
- Peifer, M. & Polakis, P. Wnt signaling in oncogenesis and embryogenesis—a look outside the nucleus. *Science* **287**, 1606–1609 (2000).
- Behrens, J. *et al.* Functional interaction of an axin homolog, conductin, with β -catenin, APC, and GSK3 β . *Science* **280**, 596–599 (1998).
- Hart, M. J., de los Santos, R., Albert, I. N., Rubinfeld, B. & Polakis, P. Downregulation of β -catenin by human Axin and its association with the APC tumor suppressor, β -catenin and GSK3 β . *Curr. Biol.* **8**, 573–581 (1998).
- Neufeld, K. L. & White, R. L. Nuclear and cytoplasmic localizations of the adenomatous polyposis coli protein. *Proc. Natl Acad. Sci. USA* **94**, 3034–3039 (1997).
- Fagotto, E., Gluck, U. & Gumbiner, B. M. Nuclear localization signal-independent and importin/karyopherin-independent nuclear import of β -catenin. *Curr. Biol.* **8**, 181–190 (1998).
- Yokoyama, F., Imamoto, N., Tachibana, T. & Yoneda, Y. β -catenin can be transported into the nucleus in a Ran-unassisted manner. *Mol. Biol. Cell* **10**, 1119–1131 (1999).
- Fukuda, M. *et al.* CRM1 is responsible for intracellular transport mediated by the nuclear export signal. *Nature* **390**, 308–311 (1997).
- Fornerod, M., Ohno, M., Yoshida, M. & Mattaj, J. W. CRM1 is an export receptor for leucine-rich nuclear export signals. *Cell* **90**, 1051–1060 (1997).
- Kudo, N. *et al.* Leptomycin B inhibition of signal-mediated nuclear export by direct binding to CRM1. *Exp. Cell Res.* **242**, 540–547 (1998).
- Su, L. K., Vogelstein, B. & Kinzler, K. W. Association of the APC tumor suppressor protein with catenins. *Science* **262**, 1734–1737 (1993).
- Rubinfeld, B., Albert, I., Porfiri, E., Munemitsu, S. & Polakis, P. Loss of β -catenin regulation by the APC tumor suppressor protein correlates with loss of structure due to common somatic mutations of the gene. *Cancer Res.* **57**, 4624–4630 (1997).
- Smits, R. *et al.* APC1638T: a mouse model delineating critical domains of the adenomatous polyposis coli protein involved in tumorigenesis and development. *Genes Dev.* **13**, 1309–1321 (1999).
- Shih, I. M., Yu, J., He, T. C., Vogelstein, B. & Kinzler, K. W. The β -catenin binding domain of adenomatous polyposis coli is sufficient for tumor suppression. *Cancer Res.* **60**, 1671–1676 (2000).
- Lal, G. & Gallinger, S. Familial adenomatous polyposis. *Sem. Surg. Oncol.* **18**, 314–323 (2000).
- Rowan, A. J. *et al.* APC mutations in sporadic colorectal tumors: A mutational 'hotspot' and interdependence of the 'two hits'. *Proc. Natl Acad. Sci. USA* **97**, 3352–3357 (2000).
- Näthke, I. S., Adams, C. L., Polakis, P., Sellin, J. H. & Nelson, W. J. The adenomatous polyposis coli tumor suppressor protein localizes to plasma membrane sites involved in active cell migration. *J. Cell Biol.* **134**, 165–179 (1996).
- Hamada, F. *et al.* Identification and characterization of E-APC, a novel *Drosophila* homologue of the tumour suppressor APC. *Genes Cells* **4**, 465–474 (1999).
- Bienz, M. APC: the plot thickens. *Curr. Opin. Genet. Dev.* **9**, 595–603 (1999).
- Lantz, V. A., Clemens, S. E. & Miller, K. G. The actin cytoskeleton is required for maintenance of posterior pole plasm components in the *Drosophila* embryo. *Mech. Dev.* **85**, 111–122 (1999).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank M. Yoshida for LMB; I. Näthke for APC antibody; T. Brabletz and J. Behrens for colon cancer cells; H. Clevers for plasmids; M. West for help with the luciferase assays; C. Garvey for help with the Western blots; A. Cliffe for NES searches; and R. Arkowitz, T. Brabletz, F. Fagotto and W. Bodmer for discussion. R.R.-A. is supported by a Wellcome travelling fellowship.

Correspondence and requests for materials should be addressed to M.B. (e-mail: mb2@mrc-lmb.cam.ac.uk).

Myosin V orientates the mitotic spindle in yeast

Hongwei Yin, David Pruyne, Tim C. Huffaker & Anthony Bretscher

Department of Molecular Biology and Genetics, Biotechnology Building, Cornell University, Ithaca, New York 14853, USA

Coordination of spindle orientation with the axis of cell division is an essential process in all eukaryotes. In addition to ensuring accurate chromosomal segregation, proper spindle orientation also establishes differential cell fates and proper morphogenesis¹. In both animal and yeast cells, this process is dependent on cytoplasmic microtubules interacting with the cortical actin-based cytoskeleton^{2–5}, although the motive force was unknown. Here we show that yeast Myo2, a myosin V that translocates along polarized actin cables into the bud⁶, orientates the spindle early in the cell cycle by binding and polarizing the microtubule-associated protein Kar9 (refs 7–9). The tail domain of Myo2 that binds Kar9 also interacts with secretory vesicles¹² and vacuolar elements¹³, making it a pivotal component of yeast cell polarization.

In budding yeast, cytoplasmic microtubules originating from spindle pole bodies direct nuclear movement and orientate the spindle along the mother–bud axis to ensure faithful postmitotic inheritance of nuclei. Orientation of the spindle is established early in the cell cycle by recruitment of cytoplasmic microtubules from

one spindle pole into the bud in a process involving the actin cytoskeleton^{3,4}. Later in the cell cycle, the orientation of the spindle is maintained through an actin-insensitive mechanism, indicating that cytoplasmic microtubules become associated with cortical cues in the mother and bud that accumulate through the cell cycle³. Despite numerous studies, a molecular link between the mitotic spindle and the actin cytoskeleton has remained elusive^{5,14,15}.

The actin cytoskeleton of growing small-budded cells consists of a polarized array of actin cables and cortical actin patches focused towards the bud¹⁶. The cables are critical for orientation, as the spindle becomes misaligned in small-budded cells having a conditional and specific defect in actin cables⁵. As the class V myosin Myo2 translocates along actin cables to deliver secretory vesicles

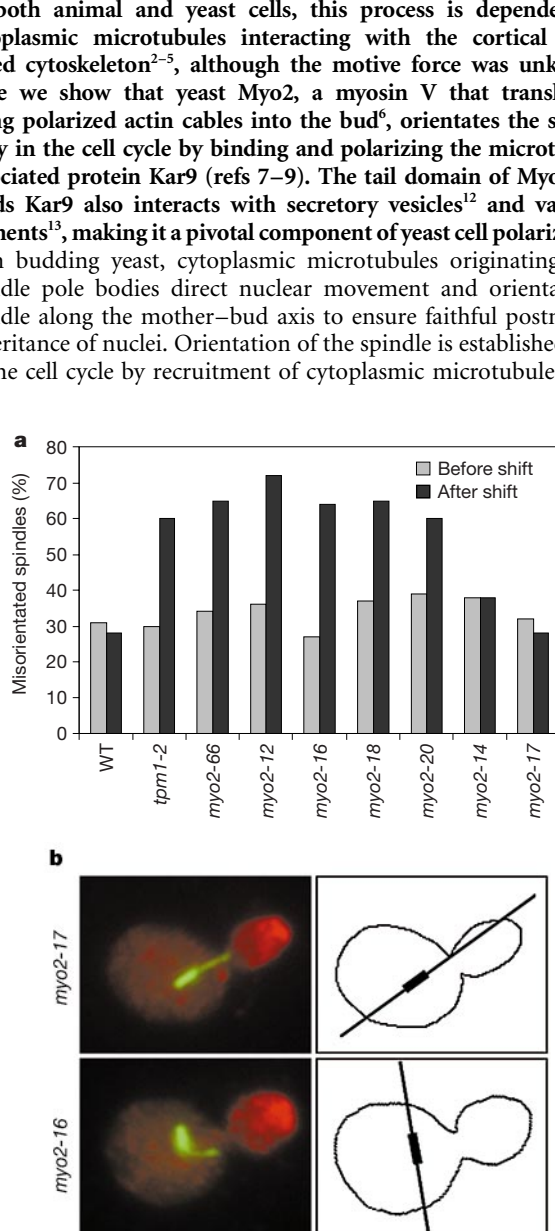


Figure 1 Spindle orientation depends on Myo2. **a**, The percentage of wild-type (WT) and conditional mutant yeast cells with misorientated spindles was determined before (grey bars) and after a 5-min shift to the restrictive temperature (black bars). **b**, Localization (left) of microtubules (green) and Myo2 (red) in temperature-shifted *myo2-17* (correctly orientated) and *myo2-16* (misorientated) cells. A spindle was considered properly orientated if an imaginary line drawn through the long axis of the spindle traversed the mother–bud neck³ (right).

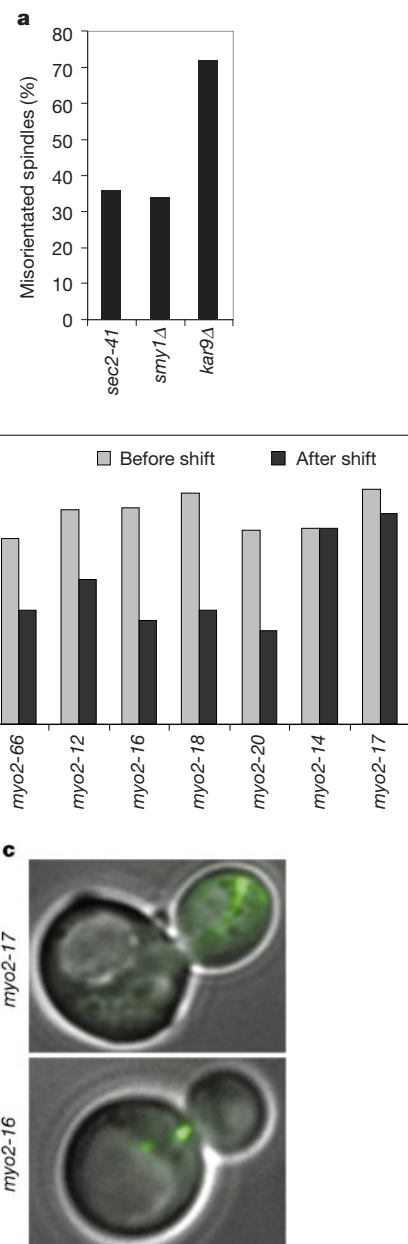


Figure 2 Kar9 polarization is affected by only those *myo2* alleles that disrupt spindle orientation. **a**, Spindle orientation in cells with a conditional defect in the secretory pathway (*sec2-41*) or loss of *SMY1* (*smv1Δ*) or *KAR9* (*kar9Δ*). **b**, Percentage of cells with GFP–Kar9 polarized to the bud cortex before (grey bars) and after a 5-min shift to the restrictive temperature (black bars). **c**, Examples of Kar9 localization in shifted *myo2-17* (polarized) and *myo2-16* (non polarized) cells.

into the bud⁶, we explored the possibility that Myo2 might also be involved in orientating the spindle.

Myo2 contains an amino-terminal motor domain and a carboxy-terminal tail domain that binds cargo. Conditional mutations specifically affecting either the motor or tail domains of Myo2 have been described^{12,17}. All of these *myo2* mutants abolish the polarized localization of secretory vesicles at their restrictive temperatures, but only the motor domain mutant fails to localize Myo2 normally to the bud¹². Wild-type and mutant yeast were shifted to the restrictive temperature for five minutes and spindle orientation determined in small- and medium-budded cells³ (Fig. 1). The temperature shift had no effect on spindle orientation in wild-type cells, whereas cells with a conditional defect in actin cables (*tpm1-2 tpm2Δ*), in the Myo2 motor domain (*myo2-66*) or in the Myo2 tail (*myo2-12*, -16, -18, -20) showed a marked increase in spindle misorientation at the restrictive temperature. Notably, only a subset of *myo2* tail mutations exhibited a spindle orientation defect; cells containing other tail mutations (*myo2-14*, -17) orientated the spindle normally, even after an hour at the restrictive temperature. The rapid orientation defect seen in the affected *myo2* alleles indicates a direct effect of the *myo2* mutations, and demonstrates the need for both motor activity and the Myo2 tail domain in spindle orientation.

Three possible schemes for the involvement of Myo2 were investigated. As Myo2 delivers secretory cargo, one possibility is that Myo2 transports some component of secretory vesicles to the bud tip that then captures and anchors the cytoplasmic microtubules there. This is unlikely because all the conditional *myo2* mutations examined affect polarized secretion, but only a subset affects spindle orientation. Moreover, spindle orientation is unaffected in a conditional *sec2* mutant (Fig. 2a) under conditions that abolish polarized targeting of secretory vesicles¹⁸.

A second possibility is that Myo2 interacts with microtubules through Smy1, a kinesin-related protein¹⁹ that colocalizes with Myo2 and binds to the Myo2 tail^{17,20}. However, this possibility was eliminated by the finding that spindle orientation was unaffected by deletion of *SMY1* (Fig. 2a).

A third possibility is that Myo2 interacts with Kar9. *kar9Δ* mutants display a karyogamy defect during mating through failure to orientate the cytoplasmic microtubules that bring the two haploid nuclei together²¹. Subsequent studies showed that strains lacking Kar9 also have a spindle orientation defect⁸ (Fig. 2a). In

small-budded cells, Kar9 is highly concentrated at the bud tip near to, or coincident with, the ends of cytoplasmic microtubules⁷, similar to the polarized localization of Myo2 (ref. 17). Although Kar9 binds microtubules through Bim1 (refs 9–11), localization of Kar9 to the bud is unaffected by depolymerization of microtubules⁷, but is disrupted by mutations and drugs that affect cortical actin organization^{8,22}. We used Kar9 tagged with green fluorescent protein (GFP–Kar9), which colocalizes with endogenous Kar9 (ref. 8), as a tracer for Kar9 distribution in the *myo2* mutants. Strikingly, there is a conditional reduction of Kar9 polarization in only those *myo2* mutants that show a spindle orientation defect, and not those showing normal spindle orientation (Fig. 2b). Again, the rapidity of this effect indicates a direct involvement of Myo2 in the polarized distribution of Kar9.

We therefore explored the possibility that Kar9 interacts with Myo2 using the yeast two-hybrid assay (Fig. 3a). Myo2 consists of an N-terminal motor domain, followed by a central domain that contains IQ motifs and a coiled-coil region, and terminates in ~500-residue tail domain. Kar9 interacts with the tail domain of wild-type Myo2, but not with the motor or central domains. Moreover, mutations of the Myo2 tail that disrupt spindle orientation abolish the interaction with Kar9. As expected, Kar9 also interacts with the microtubule-binding protein Bim1 (refs 10, 11), and the tail of Myo2 interacts with Smy1 (ref. 20). Finally, cells expressing an haemagglutinin (HA)-epitope-tagged copy of Kar9, revealed that endogenous Myo2 co-immunoprecipitates with HA–Kar9 (Fig. 3b). We conclude that the polarized distribution of Kar9 is achieved through its association with the tail of Myo2.

Yeast contains two myosin V proteins, encoded by *MYO2* and *MYO4*, both of which appear to move along polarized actin cables. The two proteins perform all the myosin-dependent polarized delivery functions known in this organism: Myo4 transports Ash1 messenger RNA into the bud of haploid cells to suppress mating type switching^{23,24}; and Myo2 delivers post-Golgi secretory vesicles¹², vacuolar elements¹³ and now Kar9 for orientation of the spindle. All the delivery functions of Myo2 involve interactions of the tail domain with different regions needed for secretory vesicle and vacuolar element delivery^{12,13}. However, it is not yet clear whether the Kar9 binding site in the Myo2 tail is distinct from those of other cargoes, or whether Myo2 can bind one or more cargoes simultaneously. A next step towards dissecting the roles of Myo2 would be the identification of *myo2* mutations that specifically affect spindle orientation and not secretion. Such

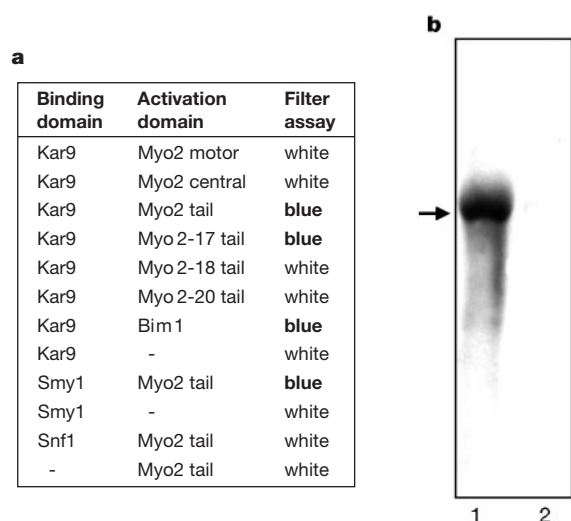


Figure 3 Kar9 interacts with the tail of Myo2. **a**, Results of two-hybrid analysis. Blue, positive interactions. Snf1 is a negative control. **b**, HA-immunoprecipitates²⁶ from *kar9Δ* cells harbouring a plasmid expressing HA-tagged Kar9 (lane 1) or untagged *KAR9* (lane 2) were resolved and blotted with Myo2-specific¹² antibodies. Arrow indicates Myo2.

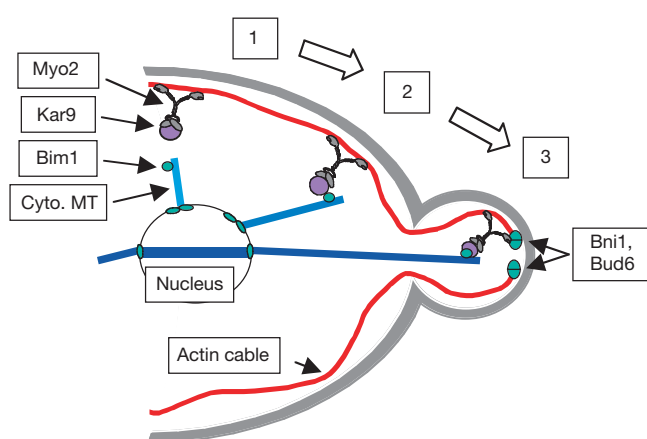


Figure 4 Working model for the establishment of spindle orientation by Myo2. Cytoplasmic microtubules (Cyto. MT) arise from spindle pole bodies early in the cell cycle. Kar9 bound to the tail of Myo2 interacts with Bim1 bound near the ends of these microtubules^{10,11,27}. Polarized actin cables, organized by cap components such as Bni1 and Bud6 (ref. 28), provide tracks for Myo2 to move to the bud tip (steps 1–3). Myosin-dependent transport, possibly in conjunction with Kip3-dependent microtubule shortening^{7,29,30}, draws the spindle towards the bud.

mutants should be viable on their own but lethal in combination with a defect in the later actin-independent spindle orientation pathway^{7,10}.

We have shown here that the microtubule-based spindle is orientated by the motor activity of a myosin V translocating along polarized actin cables. As far as we are aware, this is the first example of the direct involvement of an actin-based motor in spindle orientation. The physical interaction of Kar9 with the Myo2 tail, together with the interaction of Kar9 with Bim1 (refs 10, 11) and Bim1 with tubulin⁹, provides a molecular linkage from actin cables to microtubules. Two simple models can be envisaged for spindle orientation by these protein interactions. First, Kar9 might be transported to bud tip by Myo2 and, on delivery and stabilization there, become part of a cytoplasmic microtubule capture process. Alternatively, Myo2 might bind through Kar9 to cytoplasmic microtubules and pull them to the bud tip (Fig. 4). We favour the latter model, because of the very rapid effect the Myo2 motor and tail mutations have on spindle orientation. It will be interesting to explore whether spindle orientation is a general function of myosin V motors, which have been found in all animals and plants so far examined²⁵. □

Methods

Cells and plasmids

Yeast strains carrying the conditional *myo2* (ref. 12), *tpm2Δ tpm1-2* (ref. 6) and *sec2-41* (ref. 18) alleles, or *smv1Δ* (ref. 17) and *kar9Δ* (ref. 7) have been described. The plasmids expressing GFP–Kar9 (ref. 7) and HA-tagged Kar9 were provided by R. Miller and M. Rose.

Spindle orientation

Cells were grown at 22 °C and those containing conditional-lethal alleles were shifted to their restrictive temperatures for 5 min. Microtubules²⁴ and Myo2 (ref. 12) were visualized by immunofluorescence microscopy. Spindle orientation (see Fig. 1b) was scored in small-to medium-budded cells and at least 100 cells were analysed for each data point. This scoring criterion contains information on both the spindle position and orientation. Wild-type cells score with some apparent misorientation owing to natural variation about a mean position³.

GFP–Kar9 localization

Cells containing a YCp plasmid expressing GFP–Kar9 under control of the *GAL1* promoter growing in raffinose were shifted to 2% galactose for 6–8 h to induce expression of GFP–Kar9. Cells were then temperature shifted for 5 min, fixed and viewed for GFP fluorescence⁸.

Protein interaction analysis

For two-hybrid analysis²⁶, the Gal4 activation domain was fused to either Myo2-motor (residues 1–800), or Myo2 central domain (residues 776–1101), or wild-type or mutant Myo2 tail domains (residues 1083–1574) or Bim1. The Gal4 DNA binding domain was fused to either Kar9, Smy1 or Snf1 (negative control). Three *myo2* tail mutations that do not affect long-term stability of Myo2 were selected for two-hybrid analysis. Cells were assayed for interactions by incubating patches in 5-bromo-4-chloro-3-indolyl-β-D-galactoside (X-gal). Those that turned blue within 1 h were scored as positive. All negative interactions remained white throughout overnight incubation.

Received 24 March; accepted 31 May 2000.

- Rose, L. S. & Kemphues, K. J. Early patterning of the *C. elegans* embryo. *Annu. Rev. Genet.* **32**, 521–545 (1998).
- Sullivan, D. S. & Huffaker, T. C. Astral microtubules are not required for anaphase B in *Saccharomyces cerevisiae*. *J. Cell Biol.* **119**, 379–388 (1992).
- Theesfeld, C. L., Irazoqui, J. E., Bloom, K. & Lew, D. J. The role of actin in spindle orientation changes during the *Saccharomyces cerevisiae* cell cycle. *J. Cell Biol.* **146**, 1019–1032 (1999).
- Palmer, R. E., Sullivan, D. S., Huffaker, T. & Koshland, D. Role of astral microtubules and actin in spindle orientation and migration in the budding yeast *Saccharomyces cerevisiae*. *J. Cell Biol.* **119**, 583–593 (1992).
- Goode, B. L., Drubin, D. G. & Barnes, G. Functional cooperation between the microtubule and actin cytoskeletons. *Curr. Opin. Cell Biol.* **12**, 63–71 (2000).
- Pruyne, D. W., Schott, D. H. & Bretscher, A. Tropomyosin-containing actin cables direct the Myo2p-dependent polarized delivery of secretory vesicles in budding yeast. *J. Cell Biol.* **143**, 1931–1945 (1998).
- Miller, R. K. & Rose, M. D. Kar9p is a novel cortical protein required for cytoplasmic microtubule orientation in yeast. *J. Cell Biol.* **140**, 377–390 (1998).
- Miller, R. K., Matheos, D. & Rose, M. D. The cortical localization of the microtubule orientation protein, Kar9p, is dependent upon actin and proteins required for polarization. *J. Cell Biol.* **144**, 963–975 (1999).
- Schwartz, K., Richards, K. & Botstein, D. BIM1 Encodes a microtubule-binding protein in yeast. *Mol. Biol. Cell* **8**, 2677–2691 (1997).

- Lee, L. *et al.* Positioning of the mitotic spindle by a cortical-microtubule capture mechanism. *Science* **287**, 2260–2262 (2000).
- Korinek, W. S., Copeland, M. J., Chaudhuri, A. & Chant, J. Molecular linkage underlying microtubule orientation toward cortical sites in yeast. *Science* **287**, 2257–2259 (2000).
- Schott, D., Ho, J., Pruyne, D. & Bretscher, A. The COOH-terminal domain of Myo2p, a yeast myosin V, has a direct role in secretory vesicle targeting. *J. Cell Biol.* **147**, 791–808 (1999).
- Calett, N. L. & Weisman, L. S. The terminal tail region of a yeast myosin-V mediates its attachment to vacuole membranes and sites of polarized growth. *Proc. Natl Acad. Sci. USA* **95**, 14799–14804 (1998).
- Stearns, T. Motoring to the finish: kinesin and dynein work together to orient the yeast mitotic spindle. *J. Cell Biol.* **138**, 957–960 (1997).
- Heil-Chapdelaine, R. A., Adams, N. R. & Cooper, J. A. Formin' the connection between microtubules and the cell cortex. *J. Cell Biol.* **144**, 809–811 (1999).
- Adams, A. E. & Pringle, J. R. Relationship of actin and tubulin distribution to bud growth in wild-type and morphogenetic-mutant *Saccharomyces cerevisiae*. *J. Cell Biol.* **98**, 934–945 (1984).
- Lille, S. H. & Brown, S. S. Immunofluorescence localization of the unconventional myosin, Myo2p, and the putative kinesin-related protein, Smy1p, to the same regions of polarized growth in *Saccharomyces cerevisiae*. *J. Cell Biol.* **125**, 825–842 (1994).
- Walch-Solimena, C., Collins, R. N. & Novick, P. J. Sec2p mediates nucleotide exchange on Sec4p and is involved in polarized delivery of post-Golgi vesicles. *J. Cell Biol.* **137**, 1495–1509 (1997).
- Lillie, S. H. & Brown, S. S. Suppression of a myosin defect by a kinesin-related gene. *Nature* **356**, 358–361 (1992).
- Beningo, K. A., Lillie, S. H. & Brown, S. S. The yeast kinesin-related protein sym1p exerts its effects on the class V myosin myo2p via a physical interaction. *Mol. Biol. Cell* **11**, 691–702 (2000).
- Kurihara, L. J., Beh, C. T., Latterich, M., Schekman, R. & Rose, M. D. Nuclear congression and membrane fusion: two distinct events in the yeast karyogamy pathway. *J. Cell Biol.* **126**, 911–923 (1994).
- Lee, L., Klee, S. K., Evangelista, M., Boone, C. & Pellman, D. Control of mitotic spindle position by the *Saccharomyces cerevisiae* formin Bni1p. *J. Cell Biol.* **144**, 947–961 (1999).
- Takizawa, P. A., Sil, A., Swedlow, J. R., Herskowitz, I. & Vale, R. D. Actin-dependent localization of an RNA encoding a cell-fate determinant in yeast. *Nature* **389**, 90–93 (1997).
- Long, R. M. *et al.* Mating type switching in yeast controlled by asymmetric localization of ASH1 mRNA. *Science* **277**, 383–387 (1997).
- Mermall, V., Post, P. L. & Mooseker, M. S. Unconventional myosins in cell movement, membrane traffic, and signal transduction. *Science* **279**, 527–533 (1998).
- Chen, X. P., Yin, H. & Huffaker, T. C. The yeast spindle pole body component Spc72p interacts with Stu2p and is required for proper microtubule assembly. *J. Cell Biol.* **141**, 1169–1179 (1998).
- Tirnauer, J. S., O'Toole, E., Berrueta, L., Bierer, B. E. & Pellman, D. Yeast Bim1p promotes the G1-specific dynamics of microtubules. *J. Cell Biol.* **145**, 993–1007 (1999).
- Pruyne, D. & Bretscher, A. Polarization of cell growth in yeast. II. The role of the cortical actin cytoskeleton. *J. Cell Sci.* **113**, 365–375 (2000).
- DeZwaan, T. M., Ellingson, E., Pellman, D. & Roof, D. M. Kinesin-related KIP3 of *Saccharomyces cerevisiae* is required for a distinct step in nuclear migration. *J. Cell Biol.* **138**, 1023–1040 (1997).
- Cottingham, F. R. & Hoyt, M. A. Mitotic spindle positioning in *Saccharomyces cerevisiae* is accomplished by antagonistically acting microtubule motor proteins. *J. Cell Biol.* **138**, 1041–1053 (1997).

Acknowledgements

We thank R. Miller, M. Rose, P. Novick, S. Lillie and D. Schott for advice and reagents. This work was supported by grants from the NIH.

Correspondence should be addressed to T.C.H. (e-mail: tch4@cornell.edu) or A.B. (e-mail: apb5@cornell.edu).

Eukaryotic polymerases ι and ζ act sequentially to bypass DNA lesions

Robert E. Johnson*, M. Todd Washington*, Lajos Haracska*, Satya Prakash & Louise Prakash

Sealy Centre for Molecular Science, University of Texas Medical Branch at Galveston, 6.104 Medical Research Building, 11th and Mechanic Streets, Galveston, Texas 77555-1061, USA

* These authors contributed equally to this work.

DNA lesions can often block DNA replication, so cells possess specialized low-fidelity, and often error-prone, DNA polymerases that can bypass such lesions and promote replication of damaged DNA¹. The *Saccharomyces cerevisiae* RAD30 and human hRAD30A encode Pol η , which bypasses a cis-syn thymine-thymine dimer efficiently and accurately^{2–7}. Here we show that a related human gene, hRAD30B⁸, encodes the DNA polymerase Pol ι , which misincorporates deoxynucleotides at a high rate. To bypass damage, Pol ι specifically incorporates deoxynucleotides

opposite highly distorting or non-instructional DNA lesions. This action is combined with that of DNA polymerase Pol ζ , which is essential for damage-induced mutagenesis, to complete the lesion bypass. Pol ζ is very inefficient in inserting deoxynucleotides

opposite DNA lesions, but readily extends from such deoxynucleotides once they have been inserted. Thus, in a new model for mutagenic bypass of DNA lesions in eukaryotes, the two DNA polymerases act sequentially: Pol ι incorporates deoxynucleotides opposite DNA lesions, and Pol ζ functions as a mispair extender.

The purified human Rad30B protein is nearly homogeneous and has the expected molecular size of 81 kilodaltons (Fig. 1a). Rad30B protein was then assayed for DNA polymerase activity on primer-template DNA substrates S-1, S-2, S-3 and S-4, in which the first template nucleotide encountered by the DNA polymerase is a C, G, T or A, respectively (Fig. 1b). The Rad30B protein exhibits DNA synthesis activity, but unlike Pol η ^{2,6}, it is unable to extend the primers to the end of the template (Fig. 1b). This activity co-eluted precisely with the Rad30B protein during purification, and we noted that no DNA polymerase activity was observed in the Rad30B mutant protein in which the highly conserved D₁₂₆ and E₁₂₇ residues present in motif III (ref. 8) have been changed to alanines (data not shown). As RAD30B encodes the ninth eukaryotic DNA polymerase to be described, it has been named Pol ι (iota).

Next, we examined the efficiency and fidelity of deoxynucleotide incorporation by Pol ι on undamaged DNA^{9,10}. Pol ι was incubated with the primer-template DNA substrates and with various concentrations of correct or incorrect deoxynucleotides. Figure 2a and b shows the pattern of incorporation of G, A, T and C opposite template A and opposite template T, respectively, and Table 1 presents the fidelity of Pol ι for deoxynucleotide incorporation opposite each of the four undamaged template bases. Opposite template A, Pol ι misincorporates A with a frequency (f_{inc}) of 3.4×10^{-4} , and it misincorporates G and C with frequencies of 2×10^{-4} and 1×10^{-6} , respectively (Table 1 and Fig. 2a). Opposite templates G and C, Pol ι is much less accurate and misincorporates deoxynucleotides with an average frequency of about 2×10^{-2} and 5×10^{-2} , respectively (Table 1). Although Pol ι inserts the correct nucleotide better than the incorrect ones opposite templates G, A and C, it exhibits an unprecedented loss in fidelity opposite template T. As indicated by the V_{max}/K_m values, Pol ι inefficiently

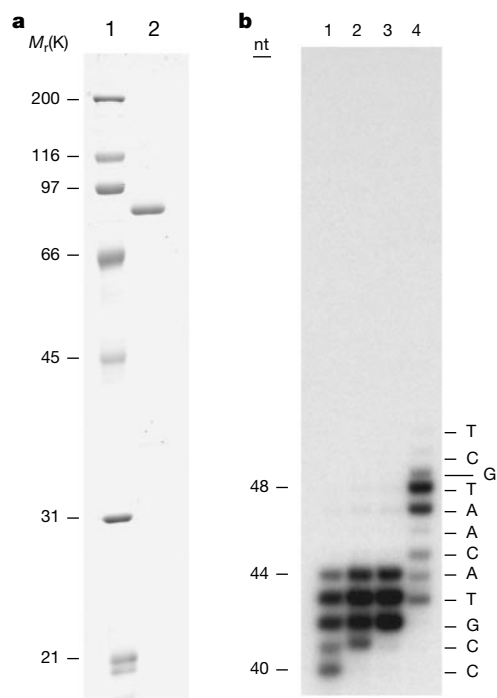


Figure 1 Purification and activity of human Rad30B protein. **a**, Purified Rad30B protein. Lane 1, molecular size markers. Lane 2, 500 ng Rad30B protein. Proteins were separated on an 11% polyacrylamide gel and stained with Coomassie Blue. **b**, DNA polymerase activity of Rad30B protein on substrates S-1, S-2, S-3 or S-4 (lanes 1–4, respectively). A portion of the template sequence is shown on the right.

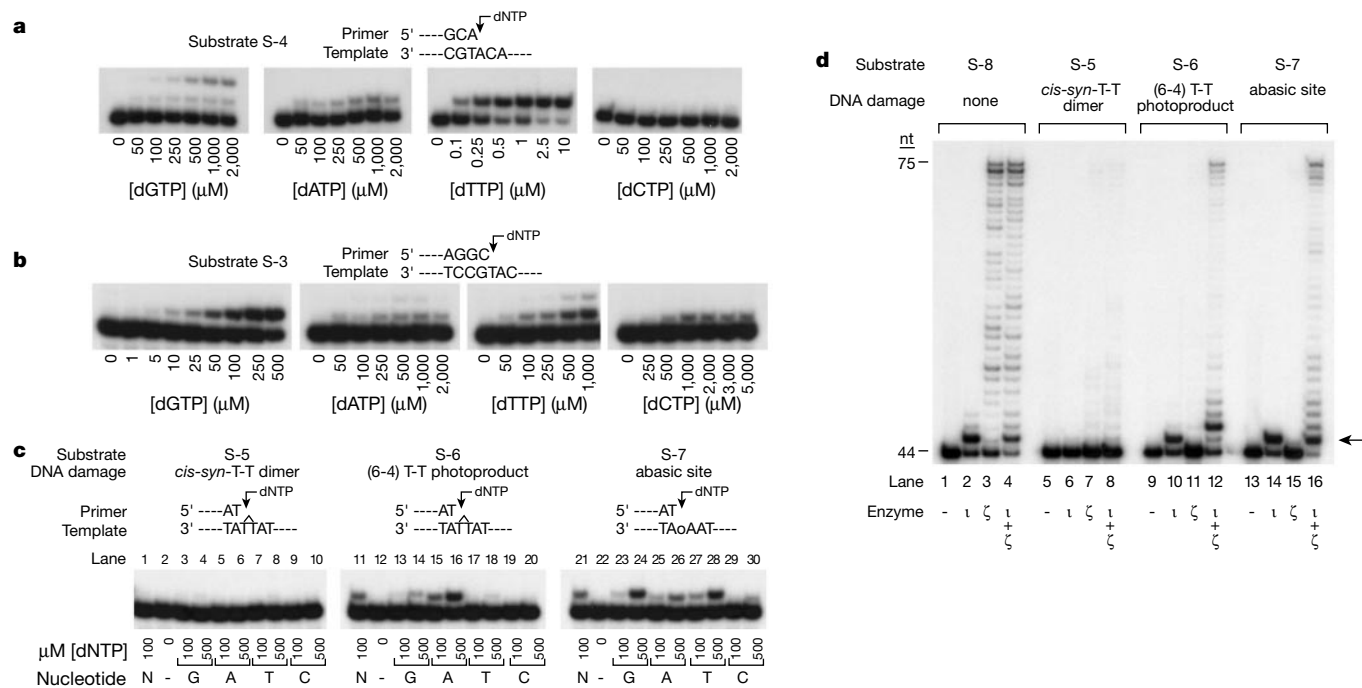


Figure 2 Deoxynucleotide incorporation by Pol ι on undamaged and damaged DNA and its role in damage bypass. **a–c**, Incorporation of deoxynucleotides opposite undamaged template A (**a**), undamaged template T (**b**), and DNA lesions (**c**). N, all four deoxynucleotides. **d**, DNA damage bypass by concerted action of Pol ι and Pol ζ . Human

DNA polymerase ι (1 nM), yeast DNA polymerase ζ (1.8 nM), or the combination of both, was incubated with DNA substrates as described in the Methods for Pol ι . Reactions were carried out at 37 °C for 10 min. Arrow indicates position of 3' T in T–T dimer or (6–4) T–T photoproduct, and position of abasic site.

incorporates A opposite T, and strikingly, it misincorporates G and T residues tenfold and fivefold better, respectively, than A (Table 1 and Fig. 2b). These observations suggest that the correct A-T primer-template base pair might be destabilized in the active site of Pol η . The much lower fidelity of Pol η at a particular subset of template bases may have arisen as a consequence of its ability to insert nucleotides opposite highly distorting DNA lesions, because an active site that is specialized for a distorted nascent base pair is no longer optimized for the normal Watson-Crick geometry.

To assess the role of Pol η in DNA-damage bypass, we examined its ability to synthesize DNA on substrates containing a *cis-syn* T-T dimer, a (6-4) T-T photoproduct, or an abasic site (Fig. 2c and d). Unlike Pol η , Pol ζ is unable to bypass a *cis-syn* T-T dimer (Fig. 2d, lane 6); moreover, it does not even incorporate a deoxynucleotide across from this lesion (Fig. 2c, lanes 1-10). However, Pol ζ is able to incorporate a deoxynucleotide opposite the 3' T of the (6-4) T-T photoproduct (Fig. 2c, lanes 11-20) and it also incorporates deoxynucleotides opposite the abasic site (Fig. 2c, lanes 21-30); but Pol ζ does not bypass either lesion (Fig. 2d, lanes 10 and 14). Next, we quantitatively examined the efficiency ($V_{\max}/K_m$) with which deoxynucleotides were incorporated opposite the 3' T of the (6-4) T-T photoproduct and opposite the abasic site (Table 1). At the 3' T of the (6-4) T-T photoproduct, Pol ζ preferentially inserts an A residue; G and T are also inserted, but they are incorporated 5-fold and 15-fold less well, respectively, than A, and C is inserted even more inefficiently (Table 1; Fig. 2c, lanes 13-20). Furthermore, the insertion of A opposite the 3' T of the (6-4) T-T photoproduct is over 30-fold more efficient than the insertion of A opposite non-damaged T (Table 1). At the abasic site, all four deoxynucleotides are inserted; the insertion of G, and T, however, is favoured over A and C (Table 1; Fig. 2c, lanes 23-30).

In yeast as well as humans, DNA polymerase ζ is essential for mutagenesis induced by DNA-damaging agents^{11,12}. We next examined whether the combination of Pol η and Pol ζ could promote lesion bypass. Whereas Pol η was unable to bypass the *cis-syn* T-T dimer, or incorporate a deoxynucleotide opposite this lesion (Fig. 2d, lane 6), Pol ζ bypasses the T-T dimer, but it does so very inefficiently (~3%) (Fig. 2d, lane 7). Pol η and Pol ζ together also do not bypass this lesion any better (Fig. 2d, lane 8). Opposite the 3' T

of the (6-4) T-T photoproduct, Pol η , but not Pol ζ , inserts a deoxynucleotide (Fig. 2d, lanes 10 and 11), and bypass of this lesion is achieved by the combined action of Pol η and Pol ζ (Fig. 2d, lane 12). Also, opposite the abasic site, Pol η , but not Pol ζ , inserts a deoxynucleotide (Fig. 2d, lanes 14 and 15), and bypass of this lesion occurs when Pol η is combined with Pol ζ (Fig. 2d, lane 16).

Steady-state kinetic analyses have further indicated that Pol ζ is very inefficient at inserting deoxynucleotides opposite the 3' T of the T-T dimer or the (6-4) T-T photoproduct, and opposite the abasic site (unpublished data). The inability of Pol ζ to efficiently insert deoxynucleotides across from DNA lesions raised the possibility that it is a high-fidelity enzyme. To test this idea, we quantitatively measured the accuracy of deoxynucleotide incorporation on undamaged DNA opposite each template base by Pol ζ . As shown in Table 2, Pol ζ incorporates deoxynucleotides with a fairly high fidelity. The f_{inc} values ranged from 4.1×10^{-3} to 1.9×10^{-5} , and in all but two cases (G incorporation opposite A, and G incorporation opposite T) the f_{inc} values were around 10^{-4} or 10^{-5} (Table 2). Eukaryotic DNA polymerase α , required for lagging-strand DNA synthesis, misincorporates nucleotides with a frequency of about 10^{-4} , and Pol δ , the eukaryotic replicative DNA polymerase, has an error frequency of about 10^{-5} (ref. 13). Thus, Pol ζ with no associated proofreading activity, has nearly the same fidelity as Pol δ , having an associated proofreading 3' to 5' exonuclease activity.

Because of its requirement in mutagenic damage bypass, the high fidelity of Pol ζ for nucleotide incorporation was unexpected. To resolve this paradox, we next examined the ability of Pol ζ to extend from mismatched primer-template termini. This was done by measuring the rate of extension from a matched or a mismatched primer terminus over a broad range of nucleotide concentrations^{9,10,14}. The efficiency of mismatch extension, f_{ext} , which is the ratio of the apparent $V_{\max}/K_m$ of extension from the mismatch to the apparent $V_{\max}/K_m$ of extension from a correct base-pair, was then determined^{9,10,14}. For all base mismatches, the f_{ext} values for Pol ζ range from approximately 10^{-1} to 10^{-2} (Table 2).

In general, DNA polymerases extend from mispaired bases at approximately the same frequency at which they insert the respective mispair^{9,14}. By contrast, avian myeloblastosis virus (AMV) reverse transcriptase extends mispairs more efficiently (10^{-1} to 10^{-5}) than it inserts mispaired bases (10^{-4} to 10^{-6})^{9,14}. Figure 3 compares the relative efficiencies of mispair extension, f_{ext} , and mispair insertion, f_{inc} , by Pol ζ for each mispair. Points lying above the dashed line indicate mispairs with a higher efficiency of extension than

Table 1 Deoxynucleotide incorporation by Pol ζ on undamaged and damaged DNA

Template residue	Incoming dNTP	$V_{\max}/K_m$ *	f_{inc}	Relative efficiency (%)
G	G	7.4×10^{-4}	4.6×10^{-3}	—
	A	1.4×10^{-3}	8.8×10^{-3}	—
	T	6.0×10^{-3}	3.8×10^{-2}	—
	C	0.16	1	—
A	G	1.7×10^{-3}	2.0×10^{-4}	—
	A	2.9×10^{-3}	3.4×10^{-4}	—
	T	8.5	1	—
	C	8.0×10^{-6} †	1.0×10^{-6}	—
T	G	8.7×10^{-3}	10	—
	A	9.0×10^{-4}	1	—
	T	4.9×10^{-3}	5.4	—
	C	2.9×10^{-4}	0.3	—
C	G	0.053	1	—
	A	2.2×10^{-3}	4.2×10^{-2}	—
	T	1.2×10^{-3}	2.3×10^{-2}	—
	C	3.7×10^{-3}	7.0×10^{-2}	—
3' T of (6-4) T-T photoproduct	G	5.5×10^{-3}	—	16
	A	0.028	—	78
	T	2×10^{-3}	—	5.5
	C	2×10^{-4} †	—	0.5
Abasic site	G	0.019	—	39
	A	9×10^{-3}	—	19
	T	0.014	—	29
	C	6.7×10^{-3}	—	13

* Units for $V_{\max}/K_m$ are nM of product per minute per μM of nucleotide. The $V_{\max}$ and K_m values and their errors are provided as Supplementary Information.

† As the nucleotide incorporation rate remained linear throughout the nucleotide concentration range used, the $V_{\max}/K_m$ value was obtained from the slope of the line.

Table 2 Deoxynucleotide incorporation and extension from matched and mismatched primer-template termini by Pol ζ on undamaged DNA

Base pair*	Incorporation		Extension†	
	$V_{\max}/K_m$ ‡	f_{inc}	$V_{\max}/K_m$ ‡	f_{ext}
G-G	2.1×10^{-3}	1.4×10^{-4}	0.21	2.4×10^{-2}
A-G	1.1×10^{-3}	7.3×10^{-5}	0.36	4.0×10^{-2}
T-G	1.7×10^{-3}	1.1×10^{-4}	0.29	3.3×10^{-2}
C-G	15	1	8.9	1
G-A	5×10^{-3}	1.3×10^{-3}	0.23	2.4×10^{-2}
A-A	9×10^{-4}	2.5×10^{-4}	0.41	4.2×10^{-2}
T-A	3.6	1	9.7	1
C-A	4×10^{-4}	1.1×10^{-4}	0.52	5.4×10^{-2}
G-T	3×10^{-2}	4.1×10^{-3}	0.51	1.0×10^{-1}
A-T	7.4	1	5.0	1
T-T	4×10^{-4}	5.4×10^{-5}	0.067	1.3×10^{-2}
C-T	1.4×10^{-4}	1.9×10^{-5}	0.52	1.0×10^{-1}
G-C	26	1	9.1	1
A-C	1.4×10^{-3}	5.4×10^{-5}	0.36	4.0×10^{-2}
T-C	1.4×10^{-3}	5.4×10^{-5}	0.068	7.5×10^{-3}
C-C	1.1×10^{-3}	4.2×10^{-5}	0.24	2.6×10^{-2}

* For incorporation of nucleotides, the first base is the incoming deoxynucleotide and the second base is in the template. For extension from a matched or mismatched primer terminus, the first base is at the 3' end of the primer and the second base is in the template opposite it.

† Extension was examined in the presence of dATP, the next correct nucleotide.

‡ Units for $V_{\max}/K_m$ are nM of product per minute per μM of nucleotide. The $V_{\max}$ and K_m values and their errors are provided as Supplementary Information.

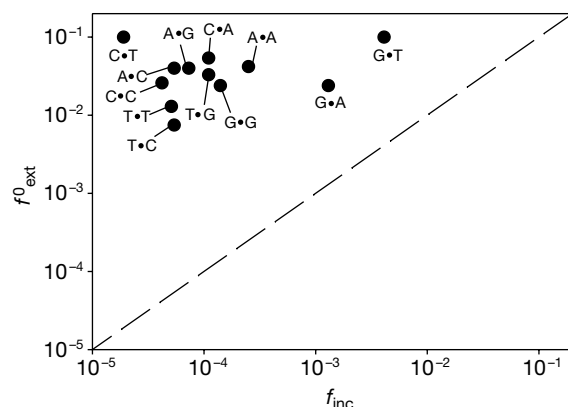


Figure 3 Comparison of Polζ mispair extension and mispair insertion efficiencies. The f_{inc} value for inserting a wrong nucleotide opposite a template base (incoming nucleotide–template base) is plotted against the corresponding f_{ext}^0 value for extending from that mispair (mispaired primer base–template base). The dashed line corresponds to $f_{ext}^0 = f_{inc}$.

insertion, and those below the line would indicate mispairs having a lower efficiency of extension than insertion. All the data points are above the line; Polζ is thus much more efficient at mispair extension than mispair formation, and for most base mispairs, f_{ext}^0 for Polζ (10^{-1} to 10^{-2}) is about three orders of magnitude greater than f_{inc} (10^{-4} to 10^{-5}). In fact, Polζ is an even more promiscuous extender of base mispairs than are the highly error prone retroviral reverse transcriptases^{9,14}.

Next, we quantitatively examined the ability of Polζ to extend from a mismatched base opposite the 3' T of the T–T dimer by measuring the incorporation of an A opposite the 5' T of the dimer. Table 3 presents the kinetic values for extension from a G, A, T or C opposite the 3' T of the T–T dimer and the non-damaged sequence. For G, T and C opposite the 3' T, the efficiency, V_{max}/K_m , of extension is about the same whether the DNA contains the T–T dimer or not. We note, however, the A opposite the 3' T of the dimer is extended tenfold less efficiently than the A opposite the non-damaged T. Furthermore, on the T–T dimer, extension of G opposite the 3' T is fourfold more efficient than extension of A. However, as Polη bypasses a *cis-syn* T–T dimer accurately^{5,6}, we assume that in the absence of Polη, mutations at the T–T dimer would occur by a T to C transition resulting from the incorporation of G opposite the 3' T of the dimer by another DNA polymerase followed by efficient extension by Polζ.

Table 3 Kinetic parameters for extension from matched and mismatched primer–template termini by Polζ on various DNAs

Base pair at the 3' primer end*	V_{max} (nM min ⁻¹)	K_m (μM)	V_{max}/K_m	f_{ext}^0
Undamaged				
G-T	1.4 ± 0.09	0.66 ± 0.02	2.1	3.8×10^{-1}
A-T	1.2 ± 0.2	0.22 ± 0.06	5.5	1
T-T	0.9 ± 0.06	1.9 ± 0.4	0.47	8.5×10^{-2}
C-T	0.95 ± 0.1	6.3 ± 1.4	0.15	2.7×10^{-2}
T–T dimer				
G-T	1.2 ± 0.04	0.55 ± 0.06	2.2	4.2
A-T	1.2 ± 0.06	2.3 ± 0.2	0.52	1
T-T	0.34 ± 0.02	1.0 ± 0.2	0.34	6.5×10^{-1}
C-T	0.26 ± 0.01	1.5 ± 0.3	0.17	3.3×10^{-1}
(6-4) T–T photoproduct				
G-T	1.2 ± 0.03	0.3 ± 0.05	4.0	2.7
A-T	0.9 ± 0.04	0.6 ± 0.1	1.5	1
T-T	0.2 ± 0.01	0.8 ± 0.2	0.3	2×10^{-1}
C-T	0.2 ± 0.01	0.4 ± 0.1	0.5	3×10^{-1}

Extension was examined in the presence of dATP, the next correct nucleotide.

*The first base is in the primer and the second base represents the 3' T of the T–T dimer or the (6-4) T–T photoproduct, or the equivalent T in undamaged DNA.

Experiments in yeast with single-stranded vectors carrying a *cis-syn* T–T dimer or a (6-4) T–T photoproduct at a unique site have indicated that whereas the *cis-syn* T–T dimer is replicated very accurately (0.4% targeted mutations), a (6-4) T–T photoproduct induces $\sim 20\%$ 3' T→C substitutions^{15,16}. Steady-state kinetic analyses of extension by Polζ from a G, A, T or C opposite the 3' T of the (6-4) photoproduct indicate that extension of G opposite the 3' T of this lesion also occurs about threefold more efficiently than extension of A (Table 3). Our finding that Polζ is most efficient at extending from G opposite the 3' T of the (6-4) T–T photoproduct is consistent with the observation that a majority of mutations in yeast at this site involve a T to C transition¹⁶. From these results, we propose that in yeast an as yet unidentified DNA polymerase preferentially inserts a G opposite the 3' T of the (6-4) T–T photoproduct, which is then extended by Polζ.

Our studies provide evidence that in humans, Polι functions in damage bypass by inserting deoxynucleotides opposite DNA lesions, whereas Polζ acts at the subsequent step of extending from them. Polι, a very low-fidelity enzyme, particularly at a subset of template bases, is unable to insert a deoxynucleotide opposite a *cis-syn* T–T dimer, which has only a modest effect on DNA structure and does not disrupt base pairing^{17,18}. Polι, however, efficiently inserts deoxynucleotides opposite a (6-4) T–T lesion, which induces a large distortion in the DNA helix¹⁹, and also opposite an abasic site, a prototypical non-instructional DNA lesion. These results suggest a specific role of Polι in damage bypass by inserting deoxynucleotides opposite highly distorting or non-instructional lesions. Polζ, on the other hand, is very inefficient in inserting deoxynucleotides opposite any of these lesions, but efficiently extends from nucleotides opposite them. Thus, Polζ functions in damage bypass by extending from nucleotides across from lesions.

Mutagenic bypass of DNA lesions in eukaryotes thus requires the sequential action of two DNA polymerases: one inserts the nucleotide opposite the lesion and the other extends from it. Further, our studies show that Polζ functions in error-prone replication of damaged DNA as a mispair extender and not as a mispair inserter. □

Methods

Expression and purification of human RAD30B encoded Polι and yeast Polζ

Human RAD30B complementary DNA was generated from HeLa total RNA and subsequently amplified using the Superscript One-step polymerase chain reaction after reverse transcription of RNA (RT-PCR) kit (Gibco BRL). All PCR-generated fragments were sequenced and found not to contain any mutations. The 2.1-kb BamHI fragment containing the wild-type hRAD30B gene was then cloned in frame with the glutathione S-transferase (GST) gene under the control of the *S. cerevisiae* galactose inducible phosphoglycerate kinase promoter in the overexpression plasmid pBJ760, generating plasmid pBJ799. Yeast strain BJ5464 harbouring plasmid pBJ799 was grown under conditions as described⁶.

To purify the hRAD30B encoded protein, yeast cells were lysed and protein was bound to a glutathione–Sephacrose 4B column as described⁶. GST–Rad30B protein bound to glutathione–Sephacrose was incubated for 4 h at 4°C with four units of thrombin, which cleaves the GST–Rad30B fusion protein eight amino acids amino-terminal from the first methionine of Rad30B. The cleavage reaction was stopped by the addition of PMSF to 0.25 mM, and protein was batch eluted. The cleaved protein was concentrated and re-equilibrated in 1 × KB (20 mM KPO₄, pH 7.4, 10% glycerol, 10 mM β-mercaptoethanol) using a microcon 30 (Amicon), and subsequently loaded onto a Mini-Q PC 3.2/3 column (Pharmacia). The column was washed with 10 column volumes 1 × KB before eluting the protein with a 2.4-ml 0–600 mM KCl gradient. Polι containing fractions were aliquoted and frozen at –70°C.

Polζ was purified from yeast strain Sc334 carrying the pGST–Rev3 and pRev7 plasmids using the procedure of ref. 11, except that after elution of Polζ from the glutathione–Sephacrose 4B matrix, the protein was dialysed and then loaded onto a Mini-Q PC 3.2/3 column (Pharmacia).

DNA polymerase assays

The standard DNA polymerase reaction (5 μl) contained 25 mM Tris–HCl (pH 7.0), 5 mM MgCl₂, 5 mM dithiothreitol, 100 μg ml⁻¹ bovine serum albumin, 10% glycerol, 100 μM of each deoxynucleotide (dGTP, dATP, dTTP, dCTP) except where noted, 10 nM of 5' ³²P-labelled oligonucleotide primer annealed to an oligonucleotide DNA template containing or not containing a DNA lesion and 1 nM of Polι. Reactions were carried out at

37 °C for 5 min unless otherwise noted and terminated by the addition of EDTA to 50 mM. DNA products were resuspended in loading buffer (95% formamide, 0.3% cyanol blue, 0.3% bromophenol blue), heat denatured at 95 °C, and then resolved on 10% polyacrylamide gels containing 8 M urea.

DNA substrates

For DNA polymerase assays and fidelity determinations of Pol ζ , the following DNA substrates were used: S-1 (template C), S-2 (template G), S-3 (template T) and S-4 (template A) were generated by annealing the 75-mer template oligonucleotides (5'-AGCTACCATGCCTGCCCTCAAGAGTTTCGTAACATGCCTACACTGGAGTACCGGAGCATCGTCGTGACTGGGAAAC-3') to 5' ³²P-radiolabelled primer oligonucleotides N4264 (5'-GTTTTCACAGTCACGACGATGCTCCGGTACTCCAGTGTAG-3'), N4265 (5'-GTTTTCACAGTCACGACGATGCTCCGGTACTCCAGTGTAGG-3'), N4266 (5'-GTTTTCACAGTCACGACGATGCTCCGGTACTCCAGTGTAGGC-3'), and N4267 (5'-GTTTTCACAGTCACGACGATGCTCCGGTACTCCAGTGTAGGCA-3'), respectively.

For measuring the fidelity of nucleotide insertion by Pol ζ , we used the DNA substrates described⁷. For measuring mispair extension by Pol ζ , four 52-nucleotide oligomers were used as templates, and they had the sequence: 5'-TTCGTATNATGCCTACCTGAGTACCGGACATC GTCGT GACTG GGAA AC, where the underlined N is G, A, T, or C. Four 45-mers were used as primers, and they had the sequence 5'-GTTTTCACGTCACGACGATGCTCCGGTACTCCAGTGTAGGCA-3', where the underlined N is G, A, T or C.

For bypass and steady-state kinetic assays on damaged DNAs, a 44-nucleotide oligomer, N4309 (5'-GTTTTCACAGTCACGACGATGCTCCGGTACTCCAGTGTAGGCAT 3') was annealed to the 75-mer template (5'-AGCAAGTCACCAATGCTCAAGAGTTTCGTAATATGCTTACAGTGGAGTACCGGAGCATGCTCGTGACTGGGAAAC-3') containing either a *cis-syn* T-T dimer (S-5), a (6-4) T-T photoproduct (S-6), or no damage (S-8) at the underlined position. The position of the abasic site (a tetrahydrofuran moiety; Midland Co.) in substrate S-7 corresponds to the underlined C in the non-damaged 75-mer template used in substrates S-1 to S-4, and this template was annealed to the primer oligomer N4309. To measure extension of mismatches by Pol ζ opposite the 3' T of a T-T dimer or a (6-4) T-T photoproduct, four 45-nucleotide primers were used that are identical to N4309 except that they contain an additional G, A, T or C residue at the 3' end.

Steady-state kinetics assays

For Pol ζ , the standard DNA polymerase assay was used except that only a single nucleotide was added, and the concentrations of nucleotides were varied as indicated in the figures. For Pol ζ , 20 nM 5' ³²P-end-labelled primer-template DNA and 4 nM Pol ζ were incubated with various concentrations of the correct nucleotide (0 to 0.5 μ M) or incorrect nucleotide (0 to 5,000 μ M) in 25 mM NaPO₄, pH 7.0 buffer containing 5 mM MgCl₂, 5 mM dithiothreitol, 100 μ g ml⁻¹ bovine serum albumin, and 10% glycerol. Reactions were carried at 25 °C for 10 min. The apparent K_m and V_{max} parameters were obtained as described^{6,7} and the f_{inc} values were calculated using the following equation:

$$f_{inc} = (V_{max}/K_m)_{wrong}/(V_{max}/K_m)_{right}^{9,10,14}.$$

To examine the kinetics of mismatch extension by Pol ζ , we used 4 nM of Pol ζ and 20 nM of 5' ³²P-end-labelled primer-template DNA. On undamaged DNA, we varied the concentration of dATP from 0.01 to 0.5 μ M for the correct primer-template base pair and from 0.2 to 10 μ M for incorrect primer-template base pairs. For T-T dimer- or (6-4) T-T photoproduct-containing DNAs, we varied the concentration of dATP from 0.1 to 20 μ M. The intrinsic efficiency of mismatch extension, f_{ext}^0 , which is a constant that represents the efficiency of extending mismatched termini in competition with matched termini at equal DNA concentrations in the limit of zero next nucleotide, was calculated as described^{9,10,14}, using this equation: $f_{ext}^0 = (V_{max}/K_m)_{mismatch}/(V_{max}/K_m)_{matched}$. Here we assume that the binding affinities for the matched and mismatched primer-templates are similar, an assumption that is valid for previously studied DNA polymerases^{10,20,21}.

Received 16 June; accepted 19 July 2000.

- Johnson, R. E., Washington, M. T., Prakash, S. & Prakash, L. Bridging the gap: a family of novel DNA polymerases that replicate faulty DNA. *Proc. Natl Acad. Sci. USA* **96**, 12224–12226 (1999).
- Johnson, R. E., Prakash, S. & Prakash, L. Efficient bypass of a thymine-thymine dimer by yeast DNA polymerase, Pol η . *Science* **283**, 1001–1004 (1999).
- Johnson, R. E., Kondratieff, C. M., Prakash, S. & Prakash, L. *hRAD30* mutations in the variant form of xeroderma pigmentosum. *Science* **285**, 263–265 (1999).
- Masutani, C. *et al.* The *XPV* (xeroderma pigmentosum variant) gene encodes human DNA polymerase η . *Nature* **399**, 700–704 (1999).
- Washington, M. T., Johnson, R. E., Prakash, S. & Prakash, L. Accuracy of thymine-thymine dimer bypass by *Saccharomyces cerevisiae* DNA polymerase η . *Proc. Natl Acad. Sci. USA* **97**, 3094–3099 (2000).
- Johnson, R. E., Washington, M. T., Prakash, S. & Prakash, L. Fidelity of human DNA polymerase η . *J. Biol. Chem.* **275**, 7447–7450 (2000).
- Washington, M. T., Johnson, R. E., Prakash, S. & Prakash, L. Fidelity and processivity of *Saccharomyces cerevisiae* DNA polymerase η . *J. Biol. Chem.* **274**, 36835–36838 (1999).
- McDonald, J. P. *et al.* Novel human and mouse homologs of *Saccharomyces cerevisiae* DNA polymerase ζ . *Genomics* **60**, 20–30 (1999).
- Goodman, M. F., Creighton, S., Bloom, L. B. & Petruska, J. Biochemical basis of DNA replication fidelity. *Crit. Rev. Biochem. Mol. Biol.* **28**, 83–126 (1993).
- Creighton, S., Bloom, L. B. & Goodman, M. F. Gel fidelity assay measuring nucleotide misinsertion, exonucleolytic proofreading, and lesion bypass efficiencies. *Methods Enzymol.* **262**, 232–256 (1995).
- Nelson, J. R., Lawrence, C. W. & Hinkle, D. C. Thymine-thymine dimer bypass by yeast DNA polymerase ζ . *Science* **272**, 1646–1649 (1996).
- Gibbs, P. E. M., McGregor, W. G., Maher, V. M., Nisson, P. & Lawrence, C. W. A human homolog of the *Saccharomyces cerevisiae* REV3 gene, which encodes the catalytic subunit of DNA polymerase ζ . *Proc. Natl Acad. Sci. USA* **95**, 6876–6880 (1998).
- Thomas, D. C. *et al.* Fidelity of mammalian DNA replication and replicative DNA polymerases. *Biochem. J.* **30**, 11751–11759 (1991).
- Mendelman, L. V., Petruska, J. & Goodman, M. F. Base mispair extension kinetics. Comparison of DNA polymerase α and reverse transcriptase. *J. Biol. Chem.* **265**, 2338–2346 (1990).
- Gibbs, P. E. M., Kilbey, B. J., Banerjee, S. K. & Lawrence, C. W. The frequency and accuracy of replication past a thymine-thymine cyclobutane dimer are very different in *Saccharomyces cerevisiae* and *Escherichia coli*. *J. Bacteriol.* **175**, 2607–2612 (1993).
- Gibbs, P. E. M., Borden, A. & Lawrence, C. W. The T-T pyrimidine (6-4) pyrimidinone UV photoproduct is much less mutagenic in yeast than in *Escherichia coli*. *Nucleic Acids Res.* **23**, 1919–1922 (1995).
- Ciarrocchi, G. & Pedrini, A. M. Determination of pyrimidine dimer unwinding angle by measurement of DNA electrophoretic mobility. *J. Mol. Biol.* **155**, 177–183 (1982).
- Kim, J.-K., Patel, D. & Choi, B.-S. Contrasting structural impacts induced by *cis-syn* cyclobutane dimer and (6-4) adduct in DNA duplex decamers: implication in mutagenesis and repair activity. *Photochem. Photobiol.* **62**, 44–50 (1995).
- Kim, J.-K. & Choi, B.-S. The solution structure of DNA duplex-decamer containing the (6-4) photoproduct of thymidyl(3'→5')thymidine by NMR and relaxation matrix refinement. *Eur. J. Biochem.* **228**, 849–854 (1995).
- Wong, I., Patel, S. S. & Johnson, K. A. An induced-fit kinetic mechanism for DNA replication fidelity: direct measurement by single-turnover kinetics. *Biochemistry* **30**, 526–537 (1991).
- Huang, M.-M., Arnheim, N. & Goodman, M. F. Extension of base mispairs by *Taq* DNA polymerase: implications for single nucleotide discrimination in PCR. *Nucleic Acids Res.* **20**, 4567–4573 (1992).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

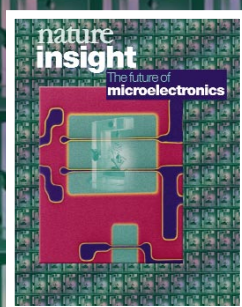
Acknowledgements

We thank D. Hinkle for plasmids pGST-Rev3 and pRev7 and R. Hodge for providing the T-T dimer- and (6-4) photoproduct-containing DNAs, the construction of which was supported by a National Institute of Environmental Health Science (NIEHS) Centre Grant. This work was supported by a grant from the NIH.

Correspondence and requests for materials should be addressed to L.P. (e-mail: lprakash@scms.utmb.edu).

nature insight

The future of microelectronics



Cover illustration

Set against repeated images of the world's first transistor, developed by Brattain and Bardeen in 1947, is a scanning electron microscopic view of phase-shifted $0.12\text{-}\mu\text{m}$ gates in one of today's state-of-the-art digital signal processors. (Images courtesy of Lucent Technologies' Bell Labs.)

The remarkable success of the semiconductor industry is well described by 'Moore's law', essentially a prescription for future progress made back in 1965, which has held to the present day: every three years will see a new generation of memory chips and microprocessors, in which the device size will reduce by 33%, the chip size will increase by 50%, and the number of components on a chip will quadruple. This has fuelled a thirst for cheaper electronic memory and increasingly powerful microprocessors that has yet to be satisfied.

But such a trend cannot continue indefinitely; indeed, it is surprising to many that it has been sustained for as long as it has. For example, if the present miniaturization trend continues, the narrowest features on microelectronic circuitry will be only a few atoms across by the end of this decade. This collection of reviews will therefore focus on some of the most pressing technological and fundamental problems that are — or will be — faced by the semiconductor industry if it is to continue to satisfy the relentless consumer demand for speed and computational power.

On page 1023, Paul Peercy provides an overview of some of the economic and technical challenges that the semiconductor industry will encounter in the next few years as it attempts to follow Moore's law. As he points out, for many of these challenges there are at present no known solutions. One of the most pressing barriers to the fabrication of ever-smaller devices stems from the process that has long been used to pattern the component devices and circuitry: optical lithography. Ito and Okazaki on page 1027 describe the scope for further enhancing the power of this technique, before discussing the alternative lithographic techniques that are available to the industry. Looking slightly farther ahead, Angus Kington and colleagues on page 1032 anticipate the need to replace silicon dioxide — a mainstay material for both memory and logic devices — with dielectrics of higher permittivity, emphasizing the magnitude of the technology shift and resources that will be required. As device dimensions approach the nanometre scale, quantum effects will have an increasingly important role in determining device properties, and may eventually necessitate the use of conceptually new device architectures. Using the single-electron transistor as an example, Devoret and Schoelkopf on page 1039 describe how these quantum effects can be harnessed to achieve new and improved device functionality. Finally, no such collection would be complete without a speculative look to the distant future. On page 1047, Seth Lloyd puts practical issues firmly to one side and investigates just how far (in terms of computational power) the laws of physics will let one go. The answer is very far indeed, if you are prepared to consider such possibilities as laptops that operate at stellar temperatures or are constructed from a black hole.

Although this collection is not intended to be comprehensive, we hope that it provides a flavour of the intellectual challenges faced by the semiconductor industry and illustrates the multidisciplinary nature of the endeavour. And by exploring issues of both a practical and fundamental nature, we hope that this collection offers something to pure and applied scientists alike.

Karl Ziemelis Physical Sciences Editor

Publisher and liaison for corporate support [Liz Allen \(e.allen@nature.com\)](mailto:Liz.Allen@nature.com)

The drive to miniaturization

Paul S. Peercy

The College of Engineering, University of Wisconsin, Madison, Wisconsin 53706-1691, USA

Following the introduction of silicon-based integrated circuitry over three decades ago, the integration density of such circuits has doubled every 12 to 18 months: this observation is known as Moore's law. For this historical trend to continue, significant challenges need to be overcome in several key technological areas. But for many of these challenges, there are at present no known solutions.

Since the invention of the integrated circuit in 1959, the semiconductor industry has improved the productivity of integrated circuits by 25–30% annually. Much of this improvement results from following Moore's law¹. Named after Gordon Moore, who observed that for silicon-based integrated circuits the number of transistors per square centimetre doubles every 12 months, Moore's law is essentially a statement of economic intent — that is, the industry intends to invest sufficient funds into research and development (R&D) and capital equipment to shrink the feature size at that rate. Throughout the 1970s and 80s, the actual doubling time was closer to 18 months, but for the past few years, the doubling time has been about 12 months. As a result of following Moore's law for the past 40 years, leading-edge companies now manufacture devices with feature sizes below 180 nm. Major technical challenges were encountered in achieving today's small feature sizes, and the number and difficulty of the technical challenges will increase with continued decrease in feature size.

The 1999 update of the Semiconductor Industry Association (SIA) *International Technology Roadmap for Semiconductors* (ITRS)² outlines the advances in technology that will be required for the semiconductor industry to continue its historical rate of improvement in productivity. By 2014, the end of the 15-year period addressed in the 1999 ITRS, complementary metal-oxide semiconductor (CMOS) field-effect transistors with gate lengths of 20 nm are projected, as are integrated circuit interconnects with linewidths of ~35 nm. This paper will highlight some of the challenges the industry will encounter as it attempts to continue the scaling required to follow Moore's law. Existing materials and technologies are approaching their physical limits, and technology breakthroughs, in terms of materials and processes, will be required as device sizes decrease significantly below 100 nm.

In addition to the technical challenges, the industry faces economic challenges. Major investment in R&D and capital

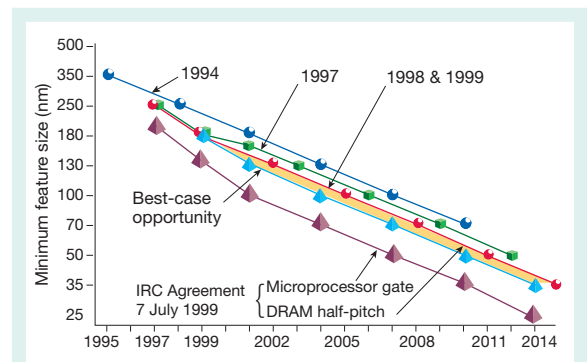


Figure 1 Comparison of the timing of the technology nodes in the 1999 ITRS with those in the 1994 and 1997 roadmaps. The first year of volume manufacturing is listed for two feature sizes, isolated gate length and DRAM half-pitch, to accommodate the replacement of memory by microprocessors as the primary technology driver.

equipment will be required if the industry is to continue its historical trend. Because suppliers have assumed responsibility for development of most of the equipment, materials and processes, and as they also control software R&D, an implicit assumption is that the supplier sector of the semiconductor industry will generate sufficient revenue to support the enormous investment in R&D required to meet the predictions of the ITRS. Technology implementation, that is, providing detailed technology solutions, is increasingly the purview of suppliers to integrated-circuit manufacturers.

Comparison of past and projected technology nodes shows that for both dynamic random-access memory (DRAM) half-pitch and gate length the semiconductor industry has scaled the feature size faster than projected in the 1994 and 1997 updates of the ITRS (Fig. 1). Scaling on this two-year cycle is projected to continue through the 100-nm gate length, before returning to the historical three-year cycle. It should be emphasized that the dates in the ITRS

Table 1 Challenges and opportunities for semiconductor R&D

Year of production:	1999	2002	2005	2008	2011	2014
DRAM half-pitch (nm)	180	130	100	70	50	35
Overlay accuracy (nm)	65	45	35	25	20	15
Microprocessor gate length (nm)	140	85–90	65	45	30–32	20–22
Critical-dimension control (nm)	14	9	6	4	3	2
Equivalent oxide thickness (nm)	1.9–2.5	1.5–1.9	1.0–1.5	0.8–1.2	0.6–0.8	0.5–0.6
Junction depth (nm)	42–70	25–43	20–33	16–26	11–19	8–13
Metal cladding (nm)	17	13	10	0	0	0
Inter-metal dielectric κ	3.5–4.0	2.7–3.5	1.6–2.2	1.5	<1.5	<1.5

Parameters set in black type, solutions exist; yellow, solutions being pursued; red, no known solutions.

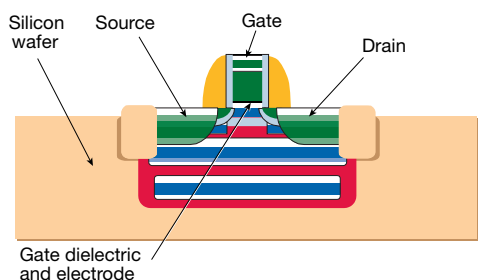


Figure 2 Schematic metal-oxide semiconductor (MOS) transistor structure, showing the source, gate and drain.

indicate the first year of volume manufacturing; suppliers must provide the equipment and processing technology well before those dates to permit device manufacturers to perform process integration and manufacturing R&D. The magnitude of the challenge is indicated in Table 1.

Other than critical-dimension control, oxide thickness and shallow-junction formation, manufacturing technology is available to extend Moore's law through 2005 to gate lengths of 65 nm and a DRAM half-pitch of 100 nm. Solutions for the subsequent technology node are unknown for three-quarters of the key technology areas. Because entries without known solutions are coloured red in the ITRS, this challenge has been labelled the 'red brick wall'. As a result of technology acceleration by the industry and the magnitude of the technical challenges, in the past two years industry has come three years closer to the 'wall'. Although the ITRS has always highlighted technology needs without known solutions, these needs were usually eight to ten years in the future and were more of an indication of how far in the future one could identify technological solutions rather than the existence of potential technical show-stoppers.

For selected technology areas I will list some of the difficult challenges that are embodied in this 'red brick wall'. Because of space limitations, my comments are limited to issues in the areas of physics and materials. Not discussed are the formidable challenges in major areas such as design, testing chips with billions of transistors operating at gigabit speeds, the process integration required to introduce a large number of new materials, or the problem of packaging integrated circuits to get the signals from those billions of transistors off the chip at gigahertz frequencies.

The CMOS transistor

Because the thickness of the gate dielectric must decrease with decreases in the gate length³, a major challenge that must be overcome in scaling the MOS transistor structure (Fig. 2) is the thickness of the silicon dioxide insulator (Fig. 3). Today's state-of-the-art CMOS transistors have a gate oxide thickness of about 2 nm. If silicon dioxide were to continue to be used for the gate insulator, the thickness is projected to decrease to 0.5 nm (about two atomic

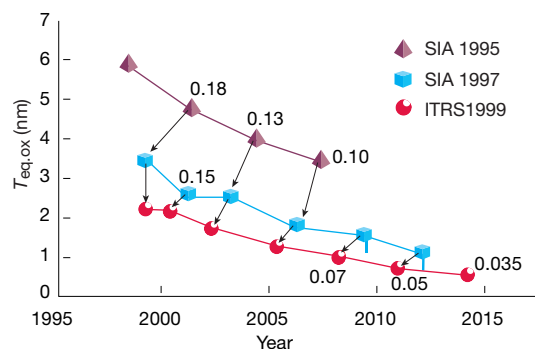


Figure 3 Projected equivalent oxide thickness ($T_{eq,ox}$) of the silicon dioxide gate dielectric from the 1999 ITRS compared with that from earlier roadmaps. The numbers in the figure refer to the gate length.

layers) by the 35-nm technology node. Tunnelling-induced leakage currents and dielectric breakdown will lead to unacceptable device performance for oxide thickness below ~1.5 nm. Research is underway with high-dielectric-constant (that is, high- κ) insulators and composite structures for gate dielectrics in the hope of replacing the thin oxides with thicker high- κ structures that have the same electrical performance as an oxide of the shown thickness. But replacement of SiO_2 would force a dramatic upheaval on the industry. With its low interface state density and low internal charge traps, SiO_2 is nature's gift to this industry and has been used since the invention of the MOS transistor. It is not clear if the projection to replace SiO_2 completely in the gate stack is realistic. The challenge of identifying and developing higher-permittivity dielectrics for replacing SiO_2 is discussed in more detail in the review article by Kingon *et al.*, pages 1032–1038.

Some of the main issues resulting from the marked reduction in effective oxide thickness are listed in Table 2, which also compares the equivalent physical oxide thickness for the 1999 ITRS with that given in the 1997 roadmap. Unsolved challenges in achieving acceptably low leakage currents and in measuring the thickness for the 1.0–1.5 nm effective oxide thickness, for both high-performance and low-power-consumption devices, appear at the 100-nm DRAM half-pitch technology node. Below 100-nm half-pitch, the structure is projected to be a dielectric sandwich containing an interfacial oxide followed by a high- κ gate dielectric. Initially, the industry expects to introduce materials such as silicon oxynitride or aluminium oxide (Al_2O_3), which have slightly a larger κ than that of SiO_2 . Subsequently, materials such as oxides of tantalum, titanium and zirconium, with values of κ in the range of 10 to 20, are projected to replace the oxynitride or aluminium oxide. Ultimately, high- κ materials such as lanthanum aluminate and zirconium titanate will replace the intermediate- κ dielectric.

If SiO_2 is replaced as the gate dielectric, to achieve adequate electrical performance, metal gate electrodes must be developed to replace today's polycrystalline silicon gate electrodes. Numerous

Table 2 Challenges resulting from reduction in gate oxide thickness

Technology node	1999 180 nm	2002 130 nm	2005 100 nm	2008 70 nm	2011 50 nm	2014 35 nm
Equivalent physical oxide thickness (nm)						
SIA 97	3.0–4.0	2.0–3.0	1.5–2.0	<1.5	<1.0	
(1999)		(2003)	(2006)	(2009)	(2012)	
ITRS 99	1.9–2.5	1.5–1.9	1.0–1.5	0.8–1.2	0.6–0.8	0.5–0.6
Gate dielectric leakage at 100 °C for HP device ($\text{nA } \mu\text{m}^{-1}$)*	5	10	20	40	80	160
Gate dielectric leakage at 100 °C for LP device ($\text{pA } \mu\text{m}^{-1}$)*	5	10	20	40	80	160

Parameters set in black type, solutions exist; yellow, solutions being pursued; red, no known solutions.

*HP device, high-performance device; LP device, low-power device.

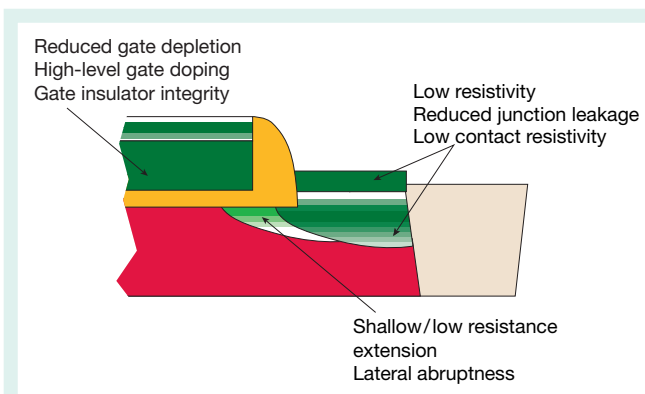


Figure 4 Illustration of doping concerns.

materials are being studied for the new gate electrodes. But as in the case of replacing SiO₂, polycrystalline silicon gates have been used by the industry since the mid 1960s, and their replacement will be far from straightforward.

With continued scaling, the junction depth of the transistors must decrease. Doping concentration and control for the ultra-shallow junctions required in the near future are approaching the limits of physics and materials. The thickness of the doped layer decreases with continued scaling, and dopant concentrations above the equilibrium solid-solubility limit will be required to provide sufficiently low resistance. Other challenges include achieving the required lateral abruptness of the junction and low junction leakage, as well as low contact resistance (Fig. 4). These properties must be achieved in a manner that ensures the integrity of the gate insulator. Ultra-low-energy ion implantation, plasma immersion doping, and a variety of rapid annealing technologies, including laser annealing, are being studied to control the dopant concentration and distribution to form ultra-shallow junctions with precise control. Successful approaches must be compatible with high- κ dielectrics and metal gate materials and provide extreme abruptness in extension/contact of the channel doping profile.

In summary, the main challenges for the future structure of transistors include critical control of the dimensions of the channel and gate length, and development of a dual-function metal-high- κ gate stack, ultra-scaled and alternate transistor structures, and silicon-compatible materials, such as SiGe or silicon-on-insulator (SOI). New materials will be required for memory cells, along with physical, chemical and electrical metrology for the devices and structures. In addition to manufacturing difficulties, there is evidence that the switching speed does not continue to increase with decreasing gate length for transistors with ultra-short gate lengths⁴. As a result, transistor structures that are based on new concepts may ultimately be required. Applications of one such structure, the single-electron transistor, are discussed by Devoret and Schoelkopf,

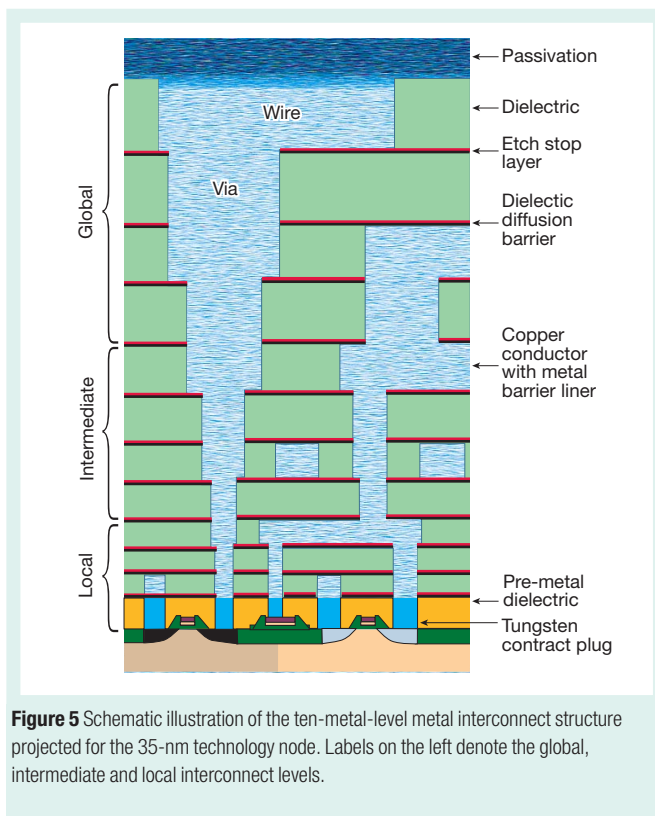


Figure 5 Schematic illustration of the ten-metal-level metal interconnect structure projected for the 35-nm technology node. Labels on the left denote the global, intermediate and local interconnect levels.

pages 1039–1046. (A more extreme extrapolation of the extent to which the laws of physics can be harnessed for the purposes of computation is presented by Lloyd, pages 1047–1054).

Interconnect

As feature sizes continue to decrease, interconnects will pose increasingly difficult challenges in terms of physics and materials, as illustrated by the cross-section of a ten-metal-level interconnect structure shown in Fig. 5. This figure shows the passivation layer, the low- κ dielectric, etch stops, dielectric diffusion barriers, copper interconnects and vias with metal barrier liners. Tungsten contact plugs through the pre-metal dielectric connect the source, gate and drain electrodes to the copper interconnects.

The increase in transistor switching speed with decreasing channel length is as important as the increased transistor density for improved performance of integrated circuits. But to benefit from this increased switching speed, signal propagation delays in interconnects between transistors must decrease with decreasing transistor switching times. Interconnect resistance increases with decreasing metal cross-section, leading to an increase in the resistance $\times$ capacitance time constant of the metal–dielectric interconnect structure and an attendant increase in the propagation

Table 3 Selected, long-term requirements of microprocessors below 100 nm

Year	2008	2011	2014
Technology node	70 nm	50 nm	35 nm
Microprocessor half-pitch	80	55	40
Microprocessor gate length (nm)	45	32	22
Number of metal levels	9	9–10	10
Barrier/cladding thickness (nm)	0	0	0
$J_{\text{max-wire}}$ (at 105 °C) (A cm ⁻²)	2.10×10^6	3.70×10^6	4.60×10^6
Effective conductor resistivity ($\mu\Omega$ cm)	1.8	<1.8	<1.8
Local wiring pitch (nm)	185	130	95
Effective dielectric constant (κ) of interlevel metal insulator	1.5	<1.5	<1.5

Parameters set in black type, solutions exist; yellow, solutions being pursued; red, no known solutions.

* $J_{\text{max-wire}}$, maximum current density in the interconnect wire.

delay. Unacceptable propagation delays with Al-SiO₂ interconnect structures caused leading-edge companies to introduce copper metals and low- κ dielectrics. The use of copper is complicated by the fact that it cannot be allowed to come in contact with the silicon wafer because copper diffuses rapidly in silicon and degrades the electrical properties.

Because distances between transistors scale with transistor feature size, local interconnect lengths will decrease with continued scaling. Thus propagation delays for local interconnects are expected to decrease with increasing transistor density and not be an issue. However, within the timescale of the current ITRS, interconnects at the global level will contribute unacceptable signal propagation delay even with Cu-low- κ interconnect structures. By introducing repeaters, the global delay can be reduced to times that are comparable to those for 0.25- μ m technology, although this solution consumes power and chip space.

Microprocessors

To better appreciate the need for materials with resistivity below that of copper and dielectrics with κ less than 1.5, consider the selected long-term requirements for microprocessors shown in Table 3.

At the 70-nm technology node (45-nm transistor gate length), microprocessors will have nine levels of metal. It is not known how to achieve the required critical current density of 4.6 MA cm⁻² in interconnects, compared with about 700 kA cm⁻² for the local wiring at today's feature size, or total current through the vias.

Some of the technology requirements for microprocessors, such as metal resistivity of 1.8 $\mu\Omega$ cm in Cu/diffusion-barrier structures, may not be achievable. One approach is to remove the barrier (giving a cladding thickness of 0). Combined with interlevel dielectrics having $\kappa < 1.5$, which implies porous structures, a barrier thickness of 0 raises significant concerns about copper diffusion. Furthermore, these dielectrics must have high physical strength and high thermal conductivity, along with good thermal stability. No solution with the required structural, mechanical, thermal, and thermal conductivity performance is known for $\kappa < 1.5$.

In addition to challenges associated with the introduction of new materials and the associated process technology, dimensional control will become increasingly difficult at smaller feature sizes, as will etching and filling structures with large aspect ratios. Perhaps more important from a fundamental perspective of materials and physics, as the metal linewidth becomes comparable to the electron mean free path, increased resistance due to scattering of electrons from the surface will become increasingly important. Specular surfaces will probably be needed to minimize the resistance of interconnects in the strong surface scattering regime, although insulators deposited on specular surfaces may not adhere sufficiently well to withstand the thermal cycles encountered in manufacturing.

Although new materials are already being introduced to address performance issues, the rate of introduction will accelerate as feature sizes are further decreased. New materials require development of process equipment and technology along with process integration. We anticipate rapid introduction of new barriers and seed layers, new low- and high- κ dielectrics, ferroelectrics, and new combinations of materials and technologies, along with new processing approaches, including *in situ* dielectric and metal formation. These materials and processes must have minimal impact on the devices. Equally important is the demonstration of adequate long-term electrical and mechanical reliability before a new material is introduced.

In addition to continued introduction of new materials, longer-term challenges include developing conceptually different interconnection approaches. The ultimate challenge is identifying and implementing solutions beyond the use of copper and low- κ dielectrics. Possible technologies include microwave interconnects with air gap insulators to achieve the required low κ and associated

low resistance $\times$ capacitance delay, or optical interconnects. So far, little research has been done on optical alternatives such as low-temperature fabrication of low-loss SiO₂ optical interconnects or on developing and integrating polymer optics. Other approaches under consideration include low-temperature operation to increase the metal conductivity. Of course, industry hopes for major research breakthroughs, such as superconductors that operate at room temperature, or new metals or polymers with a conductivity at room temperature greater than that of copper.

Lithography

The preceding discussion of transistor structure and interconnects assumed that the lithographic capability would be available to continue scaling structures to smaller feature sizes. Issues in lithography differ qualitatively from the issues in transistors and interconnects. Unlike the limits set by materials and physics that will be approached in transistors and interconnects, limits facing lithography are primarily economic. We know how to print 25-nm features — or even 10-nm features — using direct-write electron beam lithography; but we do not know how to print them cost-effectively with high throughput.

Because of the critical nature of lithography to the manufacture of all integrated circuits, several approaches for high-volume manufacture at feature sizes below 100 nm are under development. The industry has greatly extended the capability of optical lithography. Through the use of phase-shift masks, 248-nm exposure tools have been extended to the 130-nm technology node. That has delayed the introduction of 193-nm exposure tools, which, by analogy, should be extendable almost to the 70-nm node. Research is required to provide cost-effective lithography below the 70-nm node. An outline of the various lithographic options currently under investigation can be found in the article by Ito and Okazaki, pages 1027–1031.

Summary

Although we expect rapid changes in materials, materials solutions alone do not seem to be able to deliver the desired performance throughout the 15-year horizon of the 1999 ITRS. If so, it implies the end of traditional scaling. One consequence of maturation of the industry observed with the rise of the personal computer was a change in market dynamics. Acceleration of system-level solutions can be expected in an attempt to overcome the effect of scaling slowdown on productivity growth. Propagation of system-on-a-chip implementations can be expected to change the competitive picture; functionality rather than cost per unit area of silicon may become the industry driver. In anticipation of this change, optical, radio frequency, waveguide and three-dimensional approaches are being examined for added functionality.

The question facing the industry is how long can it meet the projections of the ITRS? To meet tomorrow's challenges, high-risk approaches will be combined with evolutionary approaches and new knowledge. This will require improved partnerships between the industry, universities, government laboratories and research institutes. To address the challenges, and to continue to advance performance at its historical rate, the industry will probably attempt to introduce new transistor structures. Today's approach of separately optimizing the integrated circuit, the package and the board will be rethought. It is likely to be replaced with a systems approach that combines at least the chip and the package, and perhaps the chip, package and board, leading to new structures for continued improvement in performance. □

1. Moore, G. E. Cramming more components onto integrated circuits. *Electronics* **38**, 114–116 (1965).
2. *International Technology Roadmap for Semiconductors* <<http://www.itrs.net/ntsr/publntrns.nsf>>.
3. Dennard, R. H. *et al.* Design of ion-implanted MOSFETs with very small physical dimensions. *IEEE J. Solid State Circuits* **SC-9**, 256–268 (1974).
4. Fischetti, M. V. & Laux, S. E. Performance degradation of small silicon devices caused by long-range Coulomb interactions. *Appl. Phys. Lett.* **76**, 2277–2279 (2000).

Pushing the limits of lithography

Takashi Ito* & Shinji Okazaki†

*Fujitsu Laboratories Ltd, 10 Morinosato-Wakamiya, Atsugi 243-0197, Japan

†Association of Super Advanced Electronics Technologies, 3-1 Morinosato Wakamiya, Atsugi-Shi Kanagawa 243-0198, Japan

The phenomenal rate of increase in the integration density of silicon chips has been sustained in large part by advances in optical lithography — the process that patterns and guides the fabrication of the component semiconductor devices and circuitry. Although the introduction of shorter-wavelength light sources and resolution-enhancement techniques should help maintain the current rate of device miniaturization for several more years, a point will be reached where optical lithography can no longer attain the required feature sizes. Several alternative lithographic techniques under development have the capability to overcome these resolution limits but, at present, no obvious successor to optical lithography has emerged.

The integration density of devices on a silicon chip has been increasing steadily for over three decades, driven by a combination of market forces and technological innovation. The rate of increase is approximately 50% per year in the case of dynamic random-access memories (DRAMs) and 35% per year for microprocessors. And the resulting improvements in chip performance have gone hand in hand with increases in chip yield and reduction in the cost per chip. Underlying these developments are advances in lithography — the technology by which the fine-scale patterns required for the fabrication and integration of the component devices are generated.

Lithographic requirements differ from chip to chip. The fabrication of DRAM chips has traditionally required the most advanced lithographic techniques, and the half-pitch (essentially the separation between neighbouring memory 'cells' on the chip) remains an important practical constraint in the fabrication of DRAMs. But more recently, the significant commercial pressure for increased performance and expansion of market size has resulted in the development of microprocessors, high-end logic large-scale integrated (LSI) circuits, and 'system-on-a-chip' LSI circuits that each use smaller-scale patterning than the contemporary generation of DRAMs. This is exemplified by the minimum gate length of the component transistors of microprocessors, which ultimately determines the speed with which these chips can process information. For example, the current generation of commercial microprocessors has operational speeds in excess of 500 MHz, whereas 33 MHz was the norm barely 10 years ago.

The past three decades have seen the critical length-scales of the component devices decrease by a remarkable amount, from 15 μm in the case of the first integrated circuit, to the 180 nm routinely obtained today. Yet the means by which these circuits have been patterned has, at least in a qualitative sense, remained the same: optical lithography. Can this phenomenal rate of increase in integration density be sustained indefinitely? The answer is clearly negative, in the sense that the size of the constituent atoms imposes a fundamental limit on the minimum length scale that can be ultimately attained. But even if the current rate of miniaturization continues apace, this fundamental limit remains many years off: the challenges faced by the microelectronics industry are more immediate.

The purpose of this review is therefore essentially twofold. First, through an overview of the fundamentals of optical lithography, we will show how minimum attainable

device dimensions are intimately (but not exclusively) related to the wavelength of light used, and we describe several techniques under investigation for further enhancing the resolution of this workhorse of the microelectronics industry. However, as the options available to industry are not all optical, we will also discuss the various non-optical lithographic techniques that are being explored. Figure 1 illustrates how the lithographic options vary as the critical device dimension decreases, and provides estimates of the timescales on which decisions may need to be made regarding which option(s) to adopt¹. A shift by the microelectronics industry to any non-optical lithographic technique will require the introduction of a new infrastructure of tools, materials and processing technologies, the research and development costs of which will be enormous.

Projection optical lithography

The key elements of a practical lithographic system are essentially the same for all technologies, optical and non-optical. They include the following. (1) A set of 'masks' containing the patterns of components to be fabricated in and/or on the semiconductor, the tools for making the masks, and the metrology for ensuring precise dimensions and pattern overlay. (2) An energy source (for example, a light source) for transferring the pattern from a mask to the semiconductor. (3) A medium — known as a 'photoresist' or 'resist' — for recording the pattern on the semiconductor following exposure to the source, and which allows subsequent processing of material in and/or on the underlying semiconductor. (4) Procedures for reliable detection of pattern defects, which clearly become more challenging as the critical dimensions decrease. The lithographic process — especially projection optical lithography — is closely related to the developing process in print photography, where the photographic negative has the role of the mask, and the photographic emulsion on the print is the resist.

Resolution limits

The resolution limit in conventional projection optical lithography is determined largely by the well-known Rayleigh's equation. The resolution (minimum resolvable feature) R and the corresponding depth of focus (DOF) are given by the following²:

$$R = k_1 \cdot \lambda / \text{NA}$$

$$\text{DOF} = k_2 \cdot \lambda / \text{NA}^2$$

Here λ is the exposure wavelength, NA is the numerical

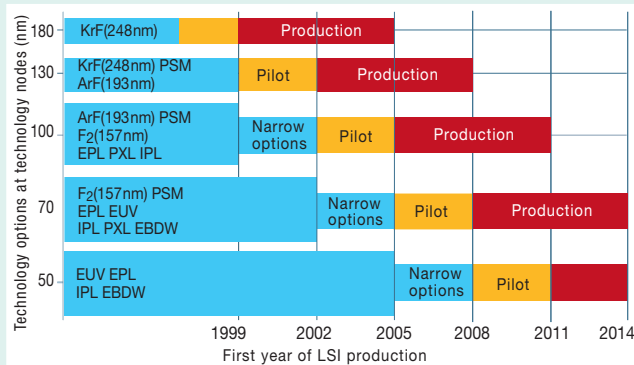


Figure 1 Technological trends in lithography. Shown are the various technologies under investigation for the development of pilot and production lines of LSI circuitry. These include: KrF (248 nm), KrF excimer laser lithography with wavelength 248 nm; ArF (193 nm), ArF excimer laser lithography with wavelength 193 nm; F₂ (157 nm), F₂ excimer laser lithography with a wavelength of 157 nm; PSM, phase-shifting mask applied to KrF, ArF and F₂; EPL, electron-projection lithography; PXL, proximity X-ray lithography; IPL, ion-projection lithography; EBDW, electron-beam direct writing.

aperture of the optical system, and k_1 and k_2 are constants that depend on the specific resist material, process technology and image-formation technique used. Figure 2 shows the evolution of projection optical lithography. It compares the trend of minimum feature size with that of the wavelength of the exposure light. When the projection system was first introduced, the required minimum feature size was relatively large compared to the wavelength of the exposure light. The numerical aperture of the lens system used was also rather small. However, as the rate of device miniaturization has proceeded faster than the rate of reduction of the exposure wavelength, projection systems with higher resolving capabilities are now required.

To obtain higher resolutions, shorter wavelength light and lens systems with larger numerical apertures can be used. In general, the minimum feature size that can be obtained is almost the same as (or slightly smaller than) the wavelength of light used for the exposure, for which one needs a relatively large numerical aperture (typically ≥ 0.5). In such high-NA lens systems, the depth of focus becomes very small, and so the exposure process becomes sensitive to slight variations in the thickness and absolute position of the resist layer³. (The smaller the depth of focus, the more rapidly a focused beam diverges on moving away from the focal point.) With the recent introduction of a 'chemical mechanical polishing' technology, the topographic variations of semiconductor surfaces have been reduced, making it possible to use extremely large NA systems; however, the margin for error becomes minute under such high-NA exposure conditions.

Resolution enhancement

To improve the resolution of an optical lithography system without introducing other impractical constraints (on, for example, wafer smoothness), several resolution-enhancement strategies have been proposed⁴⁻⁶. Figure 3 shows a schematic view of a typical projection optical-exposure system. Such a system consists of several sub-systems, and resolution-enhancement ideas can be applied at various points. These ideas generally amount to manipulating — or 'engineering' — the wavefront of the light used, and may take the form of modified illumination at the light source, phase shifting of the wavefront in the mask plane, and/or filtering with an aperture (pupil) at the projection lens. Some basic ideas of resolution enhancement are illustrated in Fig. 4, which compares image formation in the conventional system with phase shifting and modified illumination.

Consider the case when the image on the mask is a simple grating. In the conventional system (Fig. 4a), several diffracted light rays are generated by the grating. Zeroth-order light proceeds straight through the system, and several rays of higher order are generated

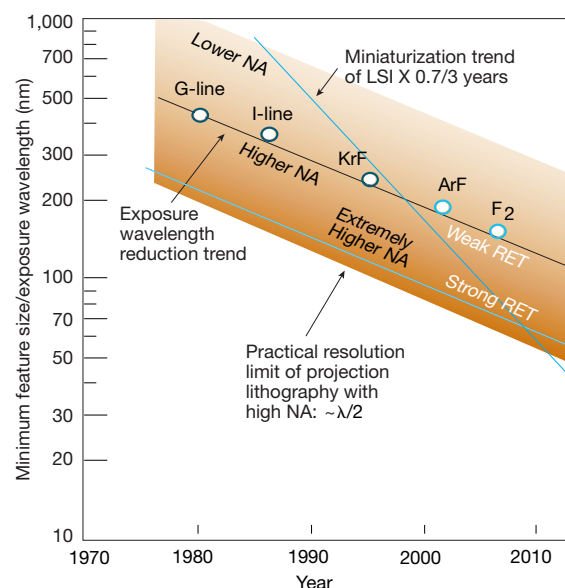


Figure 2 Comparison of trends in the wavelength of exposure light and the minimum feature size of LSI devices. The miniaturization trend is steeper than the rate at which exposure wavelengths have been reduced, illustrating the increasing practical importance of high-numerical-aperture (high-NA) systems and strong resolution-enhancement techniques (RETs). G-line and I-line refer to earlier exposure wavelengths.

with diffraction angles of θ (first order), 2θ (second order), 3θ (third order) and so on, with $\theta \approx \lambda/a$ for small θ (where a is the grating periodicity). Near the resolution limit, only the zeroth-order and first-order rays pass through the lens pupil: the transmission of at least two rays is required in order to reconstruct an image of the grating from their subsequent recombination.

Now consider the case in Fig. 4b, where the phase of the light passing through the grating mask is modified to have a periodicity twice that of the grating itself. (Phase shifting also requires a higher degree of coherency in the source light compared to the conventional case.) The transmitted wavefront is in turn modified such that the zeroth-order light is cancelled out and the diffraction angle of the first-order rays halved to 0.5θ . As a result, the spatial frequency of the patterns that can be imaged is doubled (or, put another way, the resolvable grating periodicity is halved). Although both the opaque pattern and the phase modifications on a real mask will be considerably more complex than a simple grating, this example nevertheless serves to illustrate the substantial improvement in resolution that can be achieved by this approach. Figure 5 shows resist patterns formed using KrF (248 nm) light and a phase-shift mask: structures having half the wavelength of the exposure light are clearly visible.

Finally we consider oblique (or off-axis) illumination, as shown in Fig. 4c. Zeroth-order light no longer passes through the centre of the pupil, but at an angle to the vertical. The first-order rays again emerge at angles of $\pm \theta$ with respect to the zeroth-order ray. And at the resolution limit, one of these passes through the side of the lens pupil opposite to that of the zeroth-order ray, whereas the other diffracted ray is blocked. The result is therefore geometrically equivalent to the phase-shifted case, and again leads to a doubling of the spatial frequency of the images that can be resolved.

At present, the most advanced mass-produced LSI circuits have a critical dimension of 180 nm. This is achieved by using partially coherent light from a KrF excimer laser with a wavelength of 248 nm, in combination with some resolution-enhancement techniques such as off-axis illumination, phase-shift mask, and optical proximity correction as well as advanced resist systems. To further reduce the

Resolution enhancement technologies

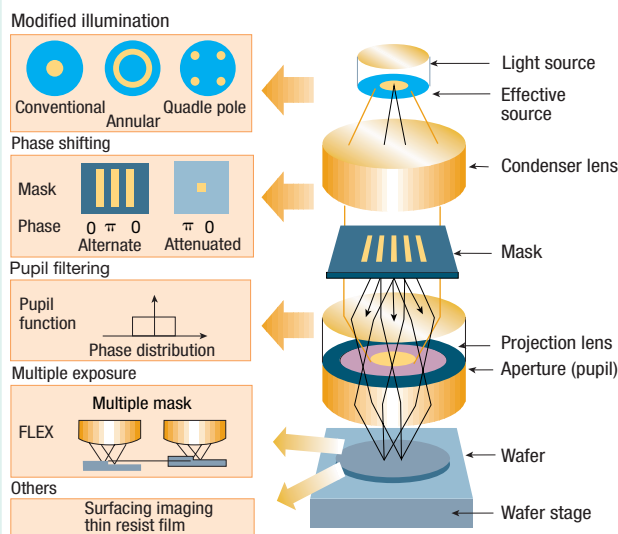


Figure 3 Schematic view of a typical exposure optical system. Resolution-enhancement techniques may be applied at various points in the optical path, as illustrated to the left of the figure.

critical dimension, wavelength reduction is being actively pursued using ArF and F₂ excimer lasers (wavelengths of 193 nm and 157 nm, respectively). The combination of shorter-wavelength light and resolution-enhancement techniques should ensure optical lithography's position as the most practical technology for LSI fabrication for at least the next 5 years.

Electron-beam lithography

Electron-beam lithography — whereby a beam of electrons rather than photons is used as the exposure source — has extremely high-resolution capabilities combined with a large depth of focus. To date, it has been used mainly in the production of masks and reticles for optical lithography. But it has also been used in the fabrication of very fine scale devices for fundamental and device-verification studies, and in the small-scale production of the specialized very high frequency devices. However, in the context of mass-produced LSI circuitry, the largest problem of electron-beam lithography is the low throughput capability of the system. Although using a finely focused electron beam makes it possible to delineate extraordinarily fine patterns, writing of chip-scale patterns with a single electron beam is a slow process. And as the complexity and finesse of mask patterns has increased, even the use of electron-beam lithography for mask writing has become prohibitively slow.

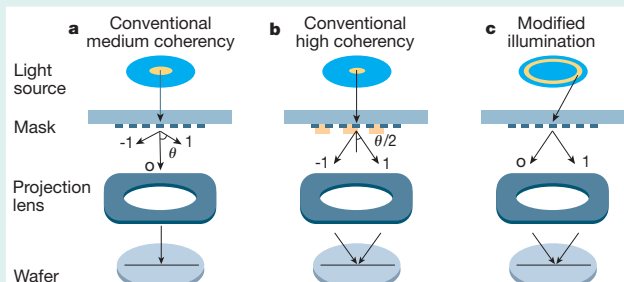


Figure 4 Grating pattern image formation using resolution enhancement. **a**, Conventional projection optics; **b**, with a phase-shifting mask; **c**, with oblique (annular) illumination.

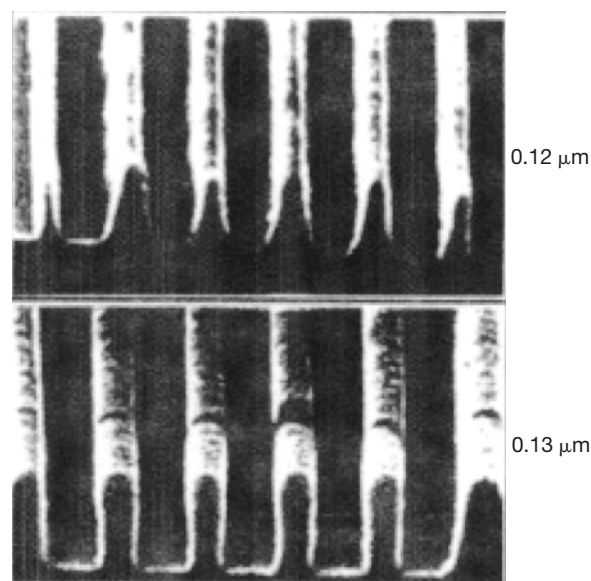


Figure 5 An example of resolution enhancement using phase-shifting technology. The exposure system uses a KrF excimer laser ($\lambda = 0.248 \mu\text{m}$; $\text{NA} = 0.55$; $\sigma = 0.3$). Structures having half the wavelength of the exposure light are clearly visible.

There has accordingly been considerable interest in developing techniques to expand the throughput of electron-beam lithography, and these have generally involved finding ways to enlarge and shape the electron beam. For example, the ability to vary the size and (rectangular) shape of the electron beam during the writing process means that the writing time is not overly sensitive to the minimum feature size: coarser features can be written with a larger beam during a single pass. But to further reduce the writing time, several areas of the pattern need to be exposed in parallel.

The cell-projection system is one such approach that makes use of parallel exposure⁷. In a conventional variable-shaped beam system, the maximum size of the electron beam is sufficient to encompass a mask containing the pattern of several memory cells (which have a repetitive structure). This can therefore be projected on the semiconductor wafer in a single exposure. Figure 6 shows a resist pattern produced by this approach: the periodically arrayed patterns at the top are produced by exposure using a cell-projection aperture, whereas the random wiring patterns at the bottom were written using a variable-shaped beam. But even with this system, throughput remains a critical issue.

A more promising approach to improve throughput is to use a projection optical system that is geometrically equivalent to that used in optical lithography: the image of an arbitrary LSI circuit is projected from a mask. Most notable of the recent attempts to use mask projection are SCALPEL (scattering angular-limited projection electron-beam lithography, developed by Bell Laboratories)⁸ and PREVAIL (projection reduction exposure with variable-axis immersion lenses, developed by IBM)⁹.

A conceptually related technique is that of ion-projection lithography (IPL), in which the electron beam is replaced by an ion beam. In principle, the much higher mass of ions should result in imaging capabilities that are less prone to distortions due to back-scattering from the substrates. But high-resolution ion-beam systems remain in a primitive state of development compared to that of electron-beam systems.

Proximity X-ray lithography

Proximity lithography is essentially a form of 'shadow printing' — a mask is held in close proximity to the semiconductor surface, and the image is simply produced on exposure by the shadow of the mask on

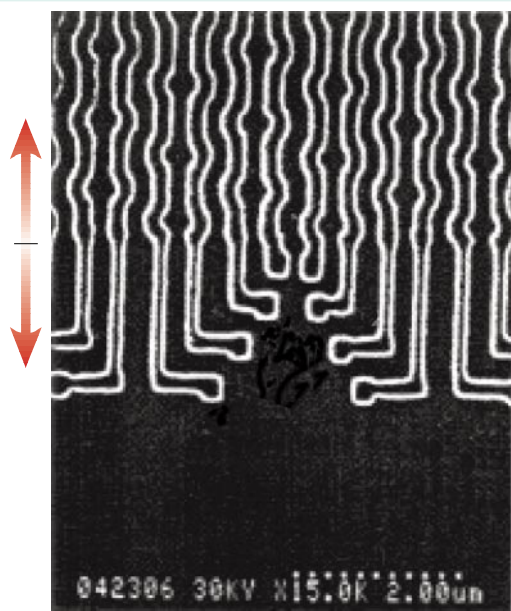


Figure 6 Resist patterns delineated by the cell-projection system. Above the indicated line are periodic patterns (memory cell patterns) exposed using cell-projection apertures; below the line are non-periodic patterns individually exposed using a variable-shaped beam (peripheral circuits patterns).

the surface. Proximity techniques were used in the early days of optical lithography, but have long since been superseded by projection techniques: the former are more susceptible to resolution-limiting diffraction effects. Resolution is less of an issue with proximity printing using X-rays, as the wavelengths used are so much smaller (~ 1 nm). In the exposure system for proximity X-ray lithography, X-rays are first collimated using a silicon carbide mirror, and are then passed through a transparent window of beryllium into a chamber containing the mask and wafer. The mask is prepared on a membrane of silicon carbide or diamond, and a layer of tantalum (patterned by direct-write electron-beam lithography) serves to absorb the X-rays and so generate the shadow on the semiconductor wafer.

Figure 7 shows a series of typical patterns produced in the resist layer following X-ray exposure and subsequent development¹⁰. The patterns shown were formed on stepped substrates, and exhibited

critical dimensions as small as 150 nm with a distribution of less than 10 nm (3σ). The final image of 100-nm node patterns demonstrates the promise of proximity X-ray lithography as a candidate for the next-generation lithographic process in LSI production. But to realize this promise, several key issues remain, such as the requirement for tight control over the mask–substrate separation (~ 15 nm) and high initial investment in equipment and infrastructures such as mask supply.

Extreme ultraviolet lithography

Extreme ultraviolet (EUV) lithography is a strong candidate for achieving critical dimensions of 70 nm and below^{11,12}. This approach uses the same principle of conventional optical-projection lithography and also obeys Rayleigh's equation shown earlier. But now the exposure wavelength is in the range 11–13 nm, which introduces its own problems.

First, as the absorption of light in this short-wavelength regime is very strong, lens-based refractive optics cannot be used in this lithographic system: an all-reflective optic system must therefore be used. A second difficulty relates to the mirrors themselves. A conventional mirror surface cannot be used at these wavelengths, so one must resort to multilayer structures that rely on interference effects to achieve reflectivity. And even then, the resulting reflectivities are rather low, typically around 60–70% at 13 nm. Accordingly, the number of mirror surfaces in the system needs to be kept as low as possible — six or fewer — to avoid significant reductions in the light intensity level.

In the case of multilayer mirror optics, the requirement of fewer optical elements means that aspherical mirrors must be introduced to achieve such capabilities. Such aspherical mirrors need to be extremely precise, with figure errors and surface roughness less than 0.1–0.2 nm. As such, an extremely high precision metrology system needs to be developed if this approach to lithography is to become viable.

The source of EUV light is another problematic issue¹⁴. At present, the best candidate is a laser-produced plasma of xenon gas. But the energy conversion ratio of laser light (from a YAG laser) to EUV light is so small that lasers of enormous power are required to achieve practical EUV exposure intensities. Different strategies, such as the use of electrical discharge lamps, are therefore being explored for providing practical light sources for EUV lithography¹⁵.

Figure 8 shows a pattern produced by an experimental EUV exposure system: 70-nm patterns are clearly delineated¹³. The insertion time for EUV lithography is thought to be 2007–2008 (see Fig. 1), leaving several years to address the practical requirements of this technology.

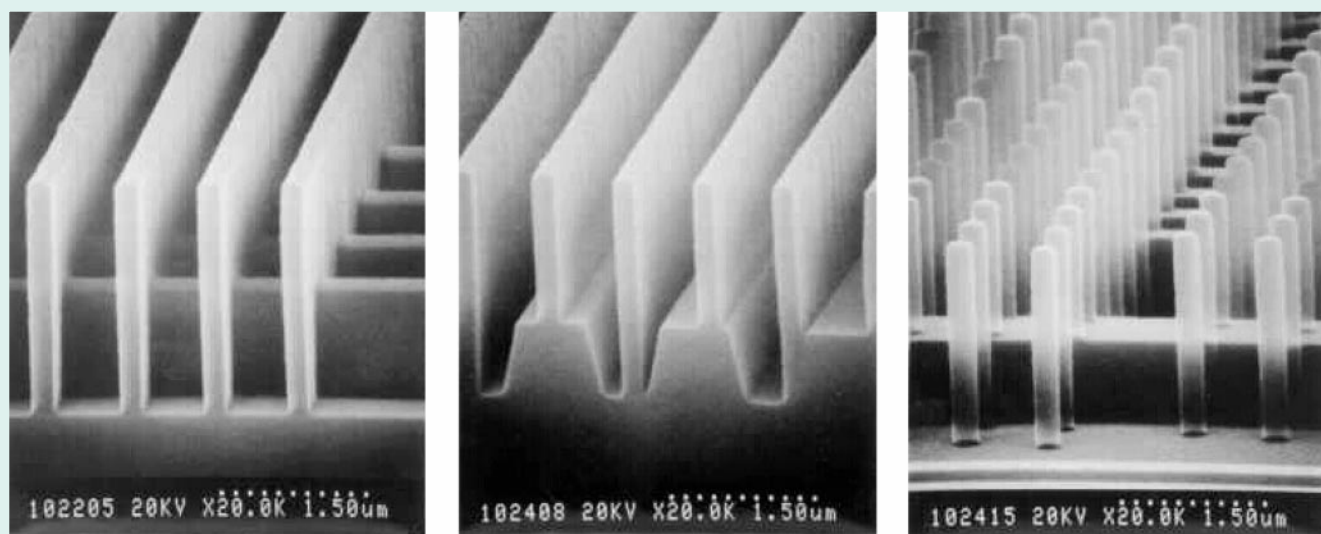


Figure 7 Resist patterns produced on different substrates using proximity X-ray lithography (PXL). (Courtesy of NTT-AT.)

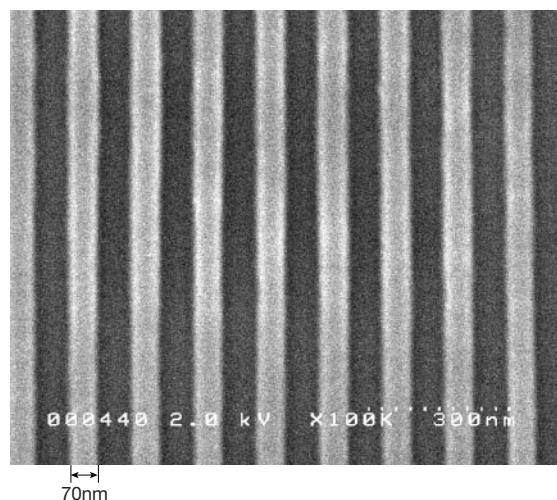


Figure 8 Seventy-nanometre line and space patterns delineated by an EUV micro-lithography system (NA = 0.15).

Other lithographies

Several other approaches have been proposed (and indeed demonstrated) for delineating patterns below 100 nm. These include nano-imprint lithography, near-field optical lithography, and direct patterning on a nanometre scale with scanning-probe microscopes. Although such approaches are useful for producing individual nano-structures for the investigation of nanometre-scale devices, the throughput is likely to always remain impracticably low for commercial application.

The 'little' picture

Optical lithography is a fundamental process in the manufacture of highly integrated microelectronic circuitry. But with the relentless commercial drive for ever smaller, faster and cheaper components, the existing technologies are rapidly being pushed to their

limits. However, optical lithography has far from reached the end of the road, and will continue to be used for some years, adopting light sources of smaller wavelengths (such as ArF and F₂ excimer lasers) and optical techniques for further enhancing resolution. Indeed, despite many years of investigation (involving the investment of hundreds of millions of dollars) into alternative lithographic techniques, no obvious replacement for optical approaches has been developed.

But the microelectronics industry is not complacent, and possible next-generation lithographic technologies using non-optical sources are still being actively explored. Whether the future of lithography lies ultimately with X-rays, electron beams or extreme ultraviolet light is far from clear at present, as critical issues remain to be solved with all alternative approaches. And as enormous costs will be incurred in developing any one technology to a stage where it becomes commercially viable, global collaboration in the microelectronics industry will be key. □

1. *International Technology Roadmap for Semiconductors* (ed. International SEMATECH) <<http://www.itrs.net/ntrs/publntrs.nsf>> (1999).
2. Klein, M. V. *Optics* (Wiley, New York, 1970).
3. Okazaki, S. Resolution limits of optical lithography. *J. Vac. Sci. Technol. B* **9**, 2829–2833 (1991).
4. Levenson, M. D. *et al.* Improving resolution in photolithography with a phase-shifting mask. *IEEE Trans. Electron Devices* **29**, 1828–1836 (1982).
5. Terasawa, T. *et al.* 0.3 μ m optical lithography using a phase shifting mask. *Proc. SPIE* **1088**, 25–33 (1989).
6. Matsuo, K. *et al.* High resolution optical lithography system using oblique incidence illumination. *Technol. Digest IEDM* **91**, 70–972 (1991).
7. Nakayama, Y. *et al.* Electron-beam cell projection lithography: a new high throughput electron direct-writing technology using a tailored Si aperture. *J. Vac. Sci. Technol. B* **8**, 1836 (1990).
8. Berger, S. D. *et al.* Projection electron beam lithography: a new approach. *J. Vac. Sci. Technol. B* **9**, 2996–2999 (1991).
9. Pfeiffer, H. C. & Stickel, W. PRIVAIL—an E-beam stepper with variable axis immersion lenses. *Microelectron. Eng.* **27**, 143–146 (1995).
10. Deguchi, K. *et al.* Patterning characteristics of a chemically-amplified negative resist in synchrotron radiation lithography. *Jpn. J. Appl. Phys.* **31**, 2954–2958 (1992).
11. Kinoshita, H. *et al.* Soft X-ray reduction lithography using multilayer mirrors. *J. Vac. Sci. Technol. B* **7**, 1648–1651 (1989).
12. Bjorkholm, J. E. *et al.* Reduction imaging at 14 nm using multilayer-coated optics: printing of features smaller than 0.1 μ m. *J. Vac. Sci. Technol. B* **8**, 1509 (1990).
13. Oizumi, H. *et al.* Sub-50 nm patterning in EUV lithography. *Jpn. J. Appl. Phys.* (in the press).
14. Kubiak, G. D. *et al.* High power extreme-ultraviolet source based on gas jets. *Proc. SPIE* **3331**, 81–89 (1998).
15. Silfvast, W. T. *et al.* High power plasma discharge source at 13.5 nm and 11.4 nm for EUV lithography. *Proc. SPIE* **3676**, 272 (1999).

Acknowledgements

A part of this work is supported by New Energy and Industrial Technology Development Organization (NEDO).

Alternative dielectrics to silicon dioxide for memory and logic devices

Angus I. Kingon*, Jon-Paul Maria* & S. K. Streiffer†

*Department of Material Science and Engineering, North Carolina State University, Raleigh, North Carolina 27695, USA (e-mail: angus_kingon@ncsu.edu)

†Materials Science Division, Argonne National Laboratory, Argonne, Illinois 60439, USA

The silicon-based microelectronics industry is rapidly approaching a point where device fabrication can no longer be simply scaled to progressively smaller sizes. Technological decisions must now be made that will substantially alter the directions along which silicon devices continue to develop. One such challenge is the need for higher permittivity dielectrics to replace silicon dioxide, the properties of which have hitherto been instrumental to the industry's success. Considerable efforts have already been made to develop replacement dielectrics for dynamic random-access memories. These developments serve to illustrate the magnitude of the now urgent problem of identifying alternatives to silicon dioxide for the gate dielectric in logic devices, such as the ubiquitous field-effect transistor.

Over the past 30 years, astounding progress has been made in silicon technology, achieved through continual scaling of semiconductor devices to ever smaller dimensions, resulting directly in a constant increase in the number of components per chip. The reduction in dimensions has gone hand-in-hand with an increase in performance and a decrease in the cost of the devices (a decrease of ~25% per year per function), fuelling an average market growth of ~15% per annum. These phenomenal trends are popularly quantified as 'Moore's law', which predicts that the number of components per chip doubles every 18 months^{1,2}; as Fig. 1 shows, Moore's law has been followed for a surprisingly long time.

Given this pace, it is interesting to note that the progress epitomized by Moore's law is best characterized as evolutionary — it has been based on a painstakingly developed and increasingly complex optimization of a relatively small set of interrelated materials and fabrication processes, but with no major revolution in fundamental device designs, and little change in the basic constituent materials (with the notable exception of the recent introduction of copper for more efficient metallization). This evolutionary nature has allowed a stable development path to be followed by the industry. Mapping of this path is undertaken on a regular basis in the *International Technology Roadmap for Semiconductors* (ITRS)³. Research and development resources have typically been allocated according to priorities established through the ITRS. However, there is currently great cause for concern within the semiconductor industry, as noted in the most recent (1999) version of the ITRS, where an increasing number of problem areas resulting from continued scaling have been identified as having 'no known solution'³.

A fundamental aspect of silicon technology is the fortuitous nature of silicon as a material — it can be reacted with oxygen or nitrogen in a controlled manner to form superb insulators with excellent mechanical, electrical and dielectric properties. These dielectrics in turn are used as core components of the two device types that represent the heart of the silicon semiconductor industry: as the capacitor dielectrics used for information storage in dynamic

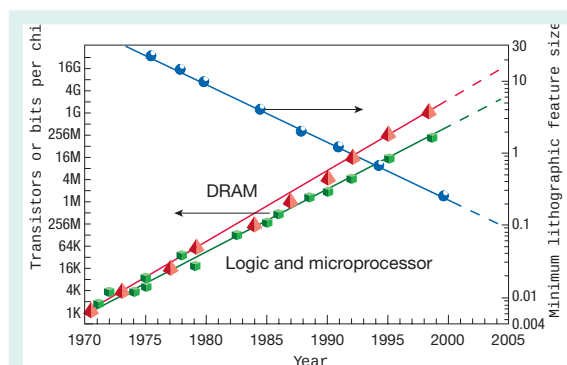


Figure 1 Schematic showing the continual increase in device density as the minimum feature size of the lithography is reduced. Device densities for both DRAM and logic devices (for example, CMOS FETs) are shown. Adapted from ref. 3 and earlier versions of the ITRS.

random-access memories (DRAMs), and as the transistor gate dielectric in complementary metal-oxide semiconductor (CMOS) field-effect transistor (FET) logic devices. In both cases, the thickness of the present dielectric, namely silicon dioxide (SiO_2) or a Si-O-N analogue, is becoming sufficiently thin that leakage currents arising from electron tunnelling through the dielectrics are posing a problem, and are viewed as a major technical barrier⁴. Although the exact thickness limits are currently the source of significant research and debate, as discussed below, one solution to the problem is the replacement of SiO_2 by an alternative insulator with a higher dielectric constant. This implies that the physical thickness of the dielectric could be increased.

In this paper we review the efforts to replace SiO_2 by higher permittivity dielectrics for the two applications. In the case of DRAMs, significant progress has been made over the past few years, and several informative reviews already exist⁵⁻⁷. In this article we point out, by example, the extent to which the new dielectrics are different to SiO_2 , thereby emphasizing the magnitude of the technology shift and resources required to adapt to the new dielectrics. The implications are then used in the discussion of the

second topic, that is, the alternative gate dielectrics for CMOS logic devices.

Alternative dielectrics for DRAM

Operation

The DRAM is the primary working medium for information storage in the ubiquitous microelectronic devices that comprise the entire litany of electronic systems. DRAM works very simply by using a submicron-sized capacitor, representing one bit of memory, to store a given amount of electrical charge: if the charge is present, it represents a digital '1', if not then the bit is a '0'. Each bit is addressed using a CMOS FET, which acts as a valve for adding or removing charge from the capacitor upon application of a voltage. It is accepted that DRAM is volatile, and information held by a DRAM is accessed at speeds approaching those at which the microprocessor operates, while archival storage is handled by other means, such as magnetic media or flash memory, that operate much more slowly. Information in a DRAM must be refreshed by rewriting each bit on regular basis, as the charge in the capacitors 'leaks' away. This refresh time is typically 256 ms.

The capacitor can be constructed rather simply by oxidizing the silicon semiconductor, and using this SiO_2 as the dielectric of the capacitor. But DRAM manufacturers face the following challenge: as chips are scaled to achieve larger device densities, the allowed capacitor area (in current devices, each cell is $\sim 0.22 \mu\text{m}^2$) becomes smaller even though the amount of stored charge remains constant^{3,5–7}. Thus, the capacitance per unit area of the wafer surface occupied by the storage node increases with each generation of device. For the early generations of capacitors (in the 1970s to the mid-1980s), the problem was simply addressed through reduction of the thickness of the SiO_2 dielectric film, as capacitance is proportional to total dielectric area, but inversely proportional to dielectric thickness. However, as the dielectric thickness is decreased, charge leakage through the dielectric becomes larger, resulting in difficulty in retaining the stored charge between refresh cycles. At small dielectric thicknesses, the leakage mechanism is quantum mechanical tunnelling, which ultimately cannot be mitigated by improvement in the quality of a given dielectric⁴. This results in a real limit to the minimum useful dielectric thickness for SiO_2 , as indicated by the simulations shown in Fig. 2.

Approach to the problem

From the mid-1980s, the DRAM manufacturers therefore concentrated on increasing the effective area of each capacitor, but without increasing its footprint in the cell. This could be achieved in two ways: by building up topographic structure on which to deposit the dielectric (Fig. 3a), or by etching a deep trench into the silicon for each memory cell, and depositing the dielectric on the walls of the trench⁵. A second lease on life was achieved by a small increase in the dielectric constant of the dielectric; the permittivity of SiO_2 (3.9) could be increased to about 6 by substitution of nitrogen into the dielectric⁸. We refer to this as the Si–O–N dielectric.

By about 1990, manufacturers became concerned that they would not be able to continue their approach of decreasing capacitor area, fearing yield and cost penalties associated with the increasing manufacturing complexity. As a result, a scientific and developmental investigation of an alternative approach coalesced around the incorporation of materials with high dielectric constants into the DRAM capacitor. Experimental studies to find direct replacements for Si–O–N focused upon amorphous or crystalline Ta_2O_5 as an intermediate permittivity material, and on the crystalline, high-permittivity, complex oxide $(\text{Ba}_x\text{Sr}_{1-x})\text{TiO}_3$, termed 'BST'. After many years of development, Ta_2O_5 -based DRAMs are now entering production at some manufacturers, and it seems that this material will be applicable for two generations of product.

BST was selected because it is a material already in use as a dielectric in bulk capacitors. Depending on composition and microstructure, the polycrystalline bulk ceramic can display a

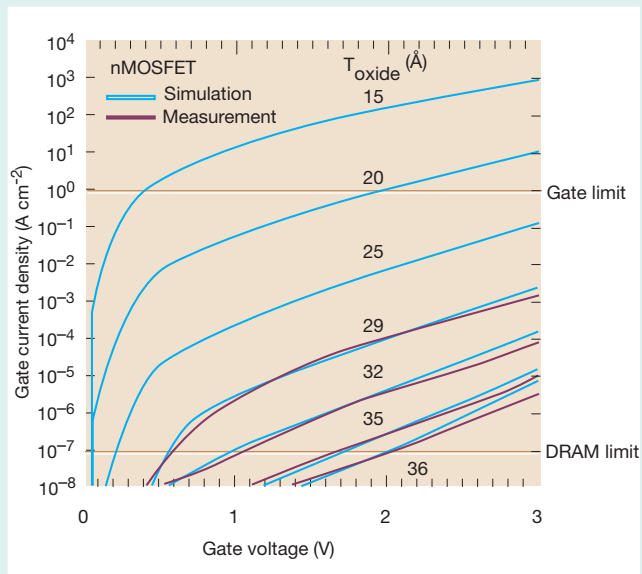


Figure 2 Plots of the current densities flowing between gate contact and channel through a SiO_2 dielectric, for various thicknesses of SiO_2 dielectric, as a function of applied voltage. Red lines are experimental, blue are calculated. Although calculated and measured for an FET, the same results are relevant for DRAMs. The horizontal lines indicate maximum allowable current densities for the two cases. Recently, experimental results have been published for oxides below 3 nm in thickness. Data for oxides as thin as 1.3 nm have been reported by Hirose *et al.*²⁴, and are in reasonable agreement with the predictions of Lo and colleagues⁴. Figure adapted from ref. 4.

relative permittivity of 1,000 to 10,000 (ref. 9); this would yield more generations of product than the relatively modest improvement in dielectric constant given by Ta_2O_5 , and thus justify a more arduous development path. It has the perovskite structure, and most studies have concentrated on compositions that were originally selected because in bulk material they yield optimum properties near room temperature. As is well documented for many materials, this assumption of generic bulk behaviour in thin films is naive, and may be used only as a rudimentary guide.

Early studies on this material as a thin-film dielectric for DRAMs originated at the DRAM manufacturers in Japan¹⁰. In 1993, the US Ultradense Capacitor Materials Processing Partnership was formed under the auspices of the Defense Advanced Research Projects Agency to further develop the material. (Members of the partnership were IBM, Micron, Texas Instruments, Advanced Technology Materials, North Carolina State University, and subsequently Varian and the University North Rhine-Westphalia.) During the period from 1993 to 1997, worldwide progress was made in developing BST in thin-film form, and demonstrating the feasibility of integrating it into DRAM cells. The capacitor schematic is shown in Fig. 3b and it is apparent that there are some differences with respect to the previous cell designs^{5–7,11}. First, the traditional SiO_2 or Si–O–N dielectrics are grown directly onto silicon (which itself is used as one capacitor electrode or plate) and covered with a metallic layer (acting as the second electrode or plate), in a configuration known as a metal–insulator–semiconductor (MIS) capacitor. But because of reaction with silicon, this could not be adopted for the BST dielectric. Therefore a configuration was required in which the BST was in contact with a different metallic bottom electrode, in a metal–insulator–metal (MIM) configuration. In addition, the BST must be processed at relatively high temperatures and oxygen partial pressures, with the implication that the electrodes must be stable against oxidation (noble metals or conductive oxides), and with the additional requirement for an oxygen barrier under the electrode to ensure that the contact (termed the plug) down to the underlying

access transistor is not oxidized. In summary, the adoption of BST results in a simpler geometry, but at the same time requires a more complex dielectric and an additional set of new materials (that is, the electrodes and barriers).

BST distinctives

In viewing the developmental progress of BST for DRAMs, it is important to realize that the task was viewed initially rather simply as the replacement of one dielectric by another. One of the important outcomes of this research and development has been the slow acceptance in the silicon community that BST displays significantly different physical behaviour than SiO_2 , as illustrated by the following four examples.

1. Dielectric phenomena. SiO_2 is a linear dielectric (permittivity is independent of field) over the field range of interest, whereas BST is strongly nonlinear (permittivity decreases with increasing field)¹². The origins of the high dielectric constant in BST are soft phonon modes (that is, coupled vibrational modes) related to its low-temperature ferroelectric phase. These may be damped out at high fields, and under particular conditions of strain inherent to growth on a silicon substrate. Strain also alters the phase transitions in the material, changing the temperature dependence of the permittivity. Although the situation is complex, satisfactory thermodynamic descriptions of thin-film dielectric behaviour of BST have been developed¹³. But in terms of DRAMs, the all-important relative dielectric permittivity of the thin films is substantially lower than that of bulk BST (for example, around 200 for a film which is 20 nm in thickness). Although this is still nearly two orders of magnitude greater than SiO_2 , it is not nearly the level of improvement originally envisaged. In general, over the thickness range between 10 nm and 1 μm , the permittivity is strongly thickness dependent^{12,13}.

2. The charge-loss mechanism. In SiO_2 -based DRAMs, the primary mechanism whereby stored charge is lost is charge leakage through the dielectric. However, in most high-permittivity dielectrics, a significant time dependence exists for the polarization response; this corresponds directly to a dielectric relaxation which is analogous to

anelasticity in polymers. This can limit the fraction of stored charge that can be read from a cell at the nanosecond timescales being approached by commercial microelectronic devices^{14–16}. This places particularly stringent tolerances on film quality¹⁷, most conveniently measured in terms of the dielectric loss tangent, which must be less than ~ 0.005 for the DRAM dielectric. Much of the BST material that has been deposited fails to meet this quality criterion, although the phenomenon is now relatively well understood.

3. Stoichiometry control. An advantage of SiO_2 is that the cation:anion stoichiometry is relatively constant over a wide range of processing conditions. The same cannot be said of BST, for which the (Ba + Sr):Ti ratio must be very tightly controlled, to about 0.2%, to achieve desired properties (Fig. 4). New deposition systems have had to be developed to achieve this control over 200-mm wafers, many wafer runs, and long time periods^{18,19}.

4. Degradation and failure. An important part of DRAM development is the ability to predict device lifetimes and failure rates. BST failure mechanisms are poorly understood and entirely different from those of SiO_2 ; thus reliability models cannot be transferred. For BST, a primary mode of failure is 'resistance degradation'^{20,21}, a phenomenon that is not observed in SiO_2 . The underlying atomistic mechanism has not been confirmed and differs subtly from that found in bulk perovskites, but is thought to involve the field-assisted transport of oxygen vacancies towards the electrodes, where they modify the interface electrical properties²⁰. Recently, a superimposed failure mechanism of time-dependent dielectric breakdown, again of different origin than similarly named phenomena in SiO_2 , has been characterized (C. B. Parker, unpublished results). This mechanism involves a statistical process that correlates with the roughness of the interface between the polycrystalline BST and the polycrystalline Pt electrode.

Although the above examples represent the substantial progress that has been made in developing a scientific understanding of BST, they emphasize the significant physical differences between BST and SiO_2 . BST has not yet been implemented as a DRAM dielectric in volume production. The target 'technology node' for implementation continues to move, from the original 256 Mbit to a current estimate of 4 Gbit or 16 Gbit. The possibility that BST will never be used is now openly discussed. Why is this? We believe that BST for DRAMs should be viewed as a revolutionary technology, rather than evolutionary, in the sense that material complexities, coupled with the different operative physical mechanisms, imply that implementation requires a new supporting technology platform (that is, tools, processes, understanding, skills and statistical data). A partial list of issues which are still being addressed includes the following: (1) materials complexity; (2) large-scale deposition tools with satisfactory stoichiometry and microstructure control; (3) yield, through all process integration steps; (4) compatibility of process atmospheres and temperatures at various steps in the process; (5) electrode and barrier requirements; (6) availability of etch equipment and processes for all materials; and (7) cost, particularly of precursor chemicals.

DRAM summary

In summary, the development of BST for DRAMs continues, in competition with alternative solutions; it is not yet clear which solution will be chosen. With hindsight, it is apparent that the implementation of BST for DRAMs is more than the simple substitution of one dielectric by another. We suggest that insufficient resources have been allocated by the industry to ensure implementation of BST in a timely fashion.

There has been, however, a particularly important indirect outcome of the research, namely, the acceptance by the semiconductor community that more complex oxides may be viable candidates for integration with silicon. This rising awareness has allowed alternative dielectrics to also be considered for the gate stack of FETs, as discussed in the following section.

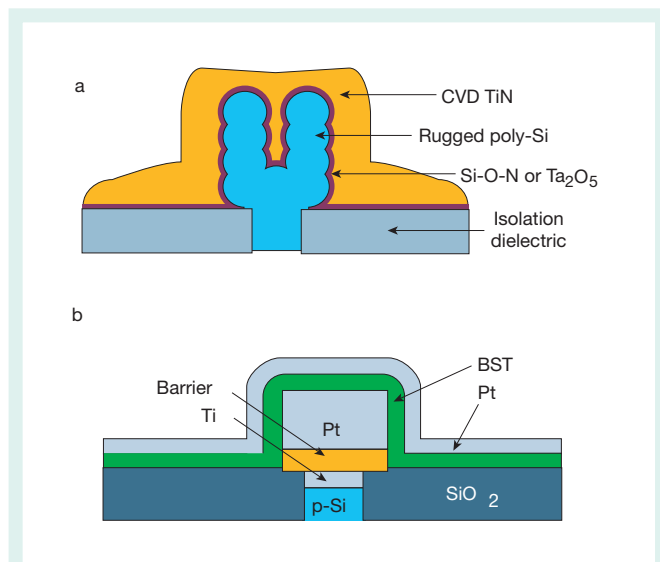


Figure 3 Schematic showing two types of DRAM architecture, adapted from ref. 11. **a**, A metal-insulator-silicon (MIS) 'stacked' cell structure which can support Si-O-N or Ta_2O_5 dielectric; **b**, schematic of the metal-insulator-metal (MIM) cell architecture under development for BST-based DRAMs. In the later case, the MIM configuration is required because an MIS configuration would lead to direct reaction between the BST and silicon. The barrier beneath the platinum bottom electrode is to minimize oxygen diffusion down to the poly-silicon contact plug that is shown. This contact plug allows direct electrical contact down to an underlying FET that acts as a switch for the capacitor.

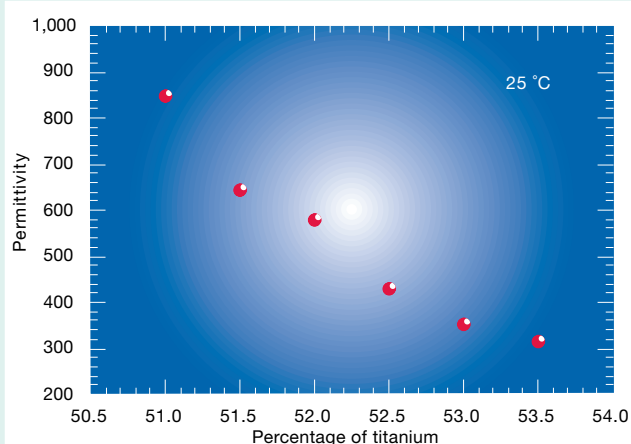


Figure 4 Plots of permittivity of $(\text{Ba}_{0.7}\text{Sr}_{0.3})\text{TiO}_3$ (BST) produced by metalorganic chemical vapour deposition (MOCVD) as a function of the $(\text{Ba} + \text{Sr})/\text{Ti}$ stoichiometry. The process is described in refs 18 and 19. The data represent the 'bulk' of the films, and have been corrected for interface capacitance effects¹². Films thickness is 40 nm. The figure emphasizes that the Ba + Sr/Ti ratio needs to be controlled tightly to minimize capacitor-to-capacitor charge-storage variations. This level of control is relatively difficult on a structure that needs to be conformally coated over the bottom electrode structure. There is also no physicochemical mechanism to self-correct the stoichiometry during the MOCVD process¹⁹. Other properties, such as dielectric losses and degradation mechanisms, are also strongly dependent on this stoichiometry. In marked contrast with bulk BST, the permittivity of the thin-film material does not display a strong dependence on the Ba:Sr ratio¹³.

Alternative gate dielectrics

The problem is closely related to that described before, in that as the device size decreases, the required capacitance per unit area is increasing beyond the capability of the present SiO_2 dielectric. In this case, however, the problem may be viewed as even more critical, in that it is central to the CMOS FET that is found in most silicon semiconductor logic and memory devices. The FET operates in a fashion that is analogous to a valve — applying a voltage between the gate electrode and the doped channel (Fig. 5) modulates the transport in the channel, between low and very high, that is, between 'on' and 'off'. The problem can be understood by looking at the cross-section (Fig. 5). Once again, scaling the FET to smaller dimensions requires simultaneous reduction in the thickness of the gate dielectric, leading to a thickness limit determined by charge tunnelling through the dielectric (Fig. 2). At high leakage current densities there are related reliability issues^{3,22}. At the same time, any scaling must maintain sufficiently high carrier mobility in the transistor channel. Considering both gate leakage and reliability, various groups have estimated that the thickness limit of SiO_2 lies somewhere between 1.0 and 1.8 nm. One source of this estimate comes from the work of Muller *et al.*²³ where Ångström-resolution electron energy-loss spectroscopy (EELS) indicated that bulk SiO_2 electronic structure is not achieved in a thermal oxide until the thickness reaches 0.7 nm. Electrically, this 'interface' material would be unsuitable for low-leakage, high-mobility device operation, implying that a capacitor

structure with two interfaces must be at least 1.4–1.5 nm thick²³. When discussing thickness limitations based on maximum allowable leakage currents, it should be noted that leakage limits are application-dependent. Estimates for maximum leakage currents in logic devices range between 1 and 10 A cm^{-2} , whereas estimates for devices in communications applications may be three orders of magnitude smaller. Recent experimental data for gate leakage as a function of SiO_2 thickness have been presented by Hirose *et al.*²⁴ on dielectrics as thin as 1.27 nm. Their data agree well with the original publication of Taur and colleagues^{4,22}. Thus, for the Muller EELS-based prediction of 1.4 nm, a leakage current in the range of tens of A cm^{-2} would be expected.

In addition to leakage current, reliability can be used to estimate the ultimate useful thinness of a SiO_2 dielectric, and Wu *et al.*²² discuss this point in detail for SiO_2 layers as thin as 1.65 nm. Moreover, they indicate that interpretation of statistical data for thin dielectrics must be treated with great care as large differences in ultimate thickness limitations can result from modest changes in the reliability parameters²².

Unfortunately, however, the ITRS indicates that the gate oxide thickness must be reduced to these estimated values for the technology generation that will begin shipping in 2005, as shown in Table 1. To meet this deadline, manufacturers need to have equipment in place by 2003 to undertake process development and optimization³. Thereafter, for following generations, the required gate oxide is significantly thinner (Table 1).

In order to meet this deadline, manufacturers are hoping that the Si–O–N dielectrics that have been successfully used in DRAMs will again be applicable, at least for one or two technology generations³. However, owing to the more stringent requirements for gate dielectrics compared to the DRAM dielectrics, it has not yet been confirmed that the Si–O–N can be manufactured with the necessary yield, and without compromising critical transistor properties such as mobility.

New dielectrics

In response to this looming crisis, manufacturers and research organizations have been studying known dielectrics with relatively high permittivities. Candidates discussed in the literature include Ta_2O_5 (refs 25, 26), TiO_2 (ref. 27) and (briefly) BST. More recently, investigations of amorphous ZrO_2 – SiO_2 and HfO_2 – SiO_2 alloys have been reported^{28,29}. The selection of the first three materials was not systematic, that is, they represent dielectrics that were known to the electronics community through other applications (for example, Ta_2O_5 is fabricated by anodic oxidation of tantalum foil for discrete capacitors). Both Ta_2O_5 and BST have become familiar to the community through the DRAM development programmes.

Considering how BST was selected as the dielectric of choice for high permittivity DRAMs, it is reasonable to state that an insufficiently broad range of factors were considered. Clearly, the dielectric should not be considered in isolation. For the gate stack it is important to ensure that the processing of the dielectric is fully compatible with the processing of the rest of the CMOS device, or in other words, that the dielectric can be integrated. At the same time, there can be no deleterious impact upon transistor performance such as channel mobility, threshold voltage and lifetime.

Selection process

As a result, a research group at North Carolina State University has developed a systematic 'dielectric selection process' in an attempt to address the gate dielectric problem. This is not described in detail in this paper, but factors that have been considered include the following: (1) chemical compatibility of candidate dielectrics with the silicon of the FET channel, under the process conditions to which the stack is exposed; (2) chemical compatibility of gate dielectric and the gate contact; (3) dielectric properties of candidate materials; (4) electronic structure of the semiconductor–dielectric interface, in particular as this structure influences leakage and transistor performance; (5) likely bulk defect structures of

Table 1 Roadmap predictions for dielectric thickness as a function of time^a

Year	1999	2002	2005	2008	2011	2014
Technology node (nm)	180	130	100	70	50	35
Gate length (nm)	140	85	65	45	32	22
Equivalent oxide thickness* (nm)	1.9–2.5	1.5–1.9	1.0–1.5	0.8–1.2	0.6–0.8	0.3–0.6

^aThe capacitance density is specified as the equivalent thickness of a SiO_2 dielectric.

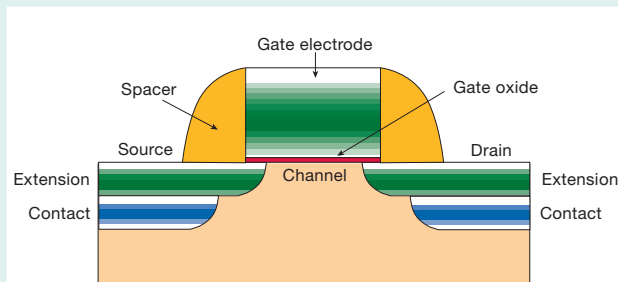


Figure 5 Schematic cross-section of a field effect transistor (FET) drawn to scale³. Note the very small thickness dimension of the gate oxide. Note also that the gate electrode (currently made of heavily doped poly-silicon) will need to be replaced at some point below the 100-nm technology node, owing to carrier-depletion effects. This replacement also promises additional complexities with regard to both physical and electrical properties. The channel corresponds to the region in which the electronic-carrier concentration is modulated by charging or discharging of the gate oxide capacitor. By modulating this concentration, the conductivity across this region (that is, between the source and drain) is changed, hence the 'valve'-type action. The source and drain regions correspond to the highly doped sections of the wafer at which two of the contacts are made (the third contact to this three-terminal device corresponds to the gate electrode). Traditionally, contacts to these regions are made from doped poly-silicon. Contacts made from alternative materials — that is, silicides or metals — may be expected in future generations of devices. With device scaling, advances are also required in source and drain formation. Currently, these conductive regions are made by implantation. However, scaled devices require that the doped high-conductivity regions have an increasingly shallow dopant profile. Laser annealing, epitaxial growth of doped material, and *in situ* doped source–drain contacts are currently under investigation.

candidate dielectrics; and (6) likely transport characteristics of the dielectric.

In the initial decision-making process, several fundamental parameters can be considered which strongly narrow the list of appropriate options. Two such parameters include the microstructure of the gate insulator, and its chemical stability at elevated temperatures in contact with silicon. These two parameters are discussed next.

Microstructures?

For gate dielectrics, three microstructures are conceivable — amorphous, polycrystalline and epitaxial. SiO_2 is used as an amorphous material, and the semiconductor industry is comfortable with this approach. Very few materials have the resistance to crystallization displayed by SiO_2 . These thin amorphous layers can be produced free from nearly all topography — one exception being crystallographic substrate steps at the silicon interface. The application of polycrystalline materials is probably limited by defects associated with grain boundaries, and by the interfacial roughness arising from potentially faceted interfaces. Interfacial roughness can increase leakage and reduce transistor mobility through charge scattering³⁰ For devices with gate lengths under 100 nm any such inhomogeneities may also result in unacceptably poor point-to-point uniformity. Epitaxial single-crystal dielectrics have also been considered³¹, primarily owing to concern about the stability of the silicon–dielectric interface. For example, oxygen that diffuses through the dielectric during or after processing can form SiO_2 at the interface. This SiO_2 is series-connected to the high-permittivity dielectric. It is straightforward to show that this has a pronounced impact, lowering the capacitance density of the dielectric stack. The epitaxial approach is therefore being considered for two reasons: the possibility of using the stability of the heteroepitaxial structure to eliminate SiO_2 formation at the interface; and the potential for achieving higher permittivities in crystalline structures. The industry is concentrating initially upon investigations of

amorphous oxides, but undertaking in parallel more fundamental research into epitaxial dielectrics that may be required for later generations.

Stability under the process conditions

The primary concern in regard to the chemical stability of gate oxides on silicon is the subsequent process conditions that the dielectric stack must tolerate after deposition. One example step from a typical process flow is the rapid thermal anneal used to achieve the appropriate dopant profile for the source and drain regions of the FET (Fig. 5). A typical anneal condition is 1,050 °C for 7 s in a reducing atmosphere. The gate dielectric must not undergo deleterious reactions with silicon under these conditions, and the amorphous candidates must not phase segregate or crystallize. If cations from the gate dielectric react or diffuse into the silicon channel, in most cases, the electrical properties will suffer. The resulting impurities are likely to scatter charge carriers, reducing their mobility. This condition requires that only chemically stable compositions are used. An alternative approach is to re-sequence the present CMOS process, undertaking the diffusion anneal before dielectric deposition. This would increase the pool of candidate dielectrics which are likely to remain amorphous, as backend process temperatures could be reduced to 500–600 °C.

It is therefore clear that knowledge of the following is required for candidate dielectrics, as a function of potential process conditions: reaction with silicon, oxygen diffusion kinetics, oxygen stoichiometry, film crystallization and component segregation. It is equally clear that this information is not readily available, and therefore the approach has been to use as 'building blocks' the simple oxides which have predicted thermodynamic stability on silicon, following the work of Hubbard and Schlom³². These simple oxides, along with their SiO_2 alloys, have received most attention: the choices include ZrO_2 , HfO_2 , Y_2O_3 and La_2O_3 . Industry is most interested in maintaining a traditional process flow, so materials that promise the best thermal stability comprise the primary choices. Currently, important data are being generated by Copel *et al.*³³, Gusev *et al.*³⁴ and E. Garfunkel *et al.* (unpublished results) using combinations of surface-sensitive ion-scattering and photoelectron spectroscopy. These groups have helped establish guidelines for high-temperature chemical reactions, oxygen diffusion, and oxide reduction. For materials such as La_2O_3 and ZrO_2 , the chemical stability in contact with silicon predicted by thermodynamics can be realized to 1,000 °C, although sensitive atmospheric control is crucial. Heat treatments to oxide films on silicon in vacuum inevitably lead to decomposition, 'freeing' of metal cations, and their subsequent diffusion into the semiconductor channel. Conversely, similar anneals in oxidizing environments lead to SiO_2 interface growth, even though the high-permittivity material can be unchanged. In effect, these sensitive measurements indicate the need for careful maintenance of particular gas atmospheres during processing, and such efforts are underway.

Another area of current research involves careful characterization of thin-film crystallization, and how this behaviour is influenced by the addition of glass-formers like SiO_2 . Much data are available in the ceramics literature, but the traditional compositions most often contain alkali or alkaline earth components which cannot be tolerated in a silicon environment. As such, a clear picture must be developed for these high-purity, two-component systems. We have studied this problem in the systems HfO_2 – SiO_2 and La_2O_3 – SiO_2 , and have generated sets of films with a composition range spanning the endmember chemistries. Diffraction methods were used to determine the onset of crystallization (to the length scales over which X-rays are sensitive). Figure 6a,b shows the trends for devitrification temperature as a function of composition. For films of similar as deposited structure, these plots indicate the minimum SiO_2 concentrations that can be allowed (and minimum dielectric dilution) under the current assumption that crystalline materials are unsuitable. For a similar fraction of SiO_2 , an advantage exists for the lanthanum-containing system in this regard.

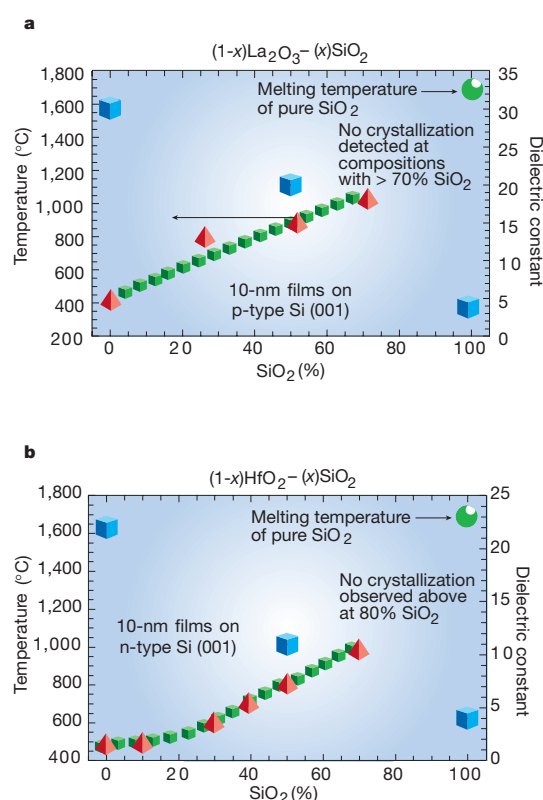


Figure 6 The compositional dependence of the crystallization temperature of **a**, $\text{La}_2\text{O}_3\text{-SiO}_2$ alloy thin films, and **b**, $\text{HfO}_2\text{-SiO}_2$ alloy thin films. In each case the film thickness is 10 nm, and the temperature exposure time is 4 min. The right axis gives the relative dielectric constants for two endmembers and an intermediate composition. It is clear that as the SiO_2 content increases, the crystallization temperatures increase, but dielectric constants fall. Data points for dielectric constant are literature bulk values.

Dielectric constants

A technical area of importance involves the ability to predict the permittivity of materials for which no reliable reference data exists. This is especially true for metal oxide–silica alloys for which almost no material has been published. It is possible to predict the dielectric constants of materials using the approach of Shannon³⁵, although this method is prone to significant errors³⁶. For systems such as $\text{HfO}_2\text{-SiO}_2$, dielectric constants of the endmember compositions are well known, although it is unclear as to which dielectric mixing laws are most appropriate for the alloy compositions. Given the type of crystallization data presented in Fig. 6, coupled with knowledge of permittivity, meaningful assessment of applicability can be made. For example, if one were to assume a linear dielectric mixing law with film composition and a process temperature requirement of 1,000 °C, the ultimate permittivities available to the hafnium- and lanthanum-containing systems would be ~7 and ~14 respectively. These dielectric measurements are currently being performed.

It is necessary to recognize the importance of carrier mobility in MOS devices. If in the process of replacing SiO_2 with an alternative material the carrier mobilities are reduced, the gains associated with a thinner dielectric will not be realized. Certainly, one of the greatest challenges associated with replacement gate oxides is to maintain the mobilities observed in Si-SiO_2 interfaces. Charge defects, interface dangling bonds and interface roughness all contribute to reduced mobility. Given the complexity and variability possible in an amorphous material, these effects are difficult to predict, and experimental verification of these effects appears to be the only approach. At the present time, there is only a limited amount of

experimental data on transistors that contain the high-permittivity dielectrics.

In the light of the issues described above, as well as the evolving experimental data, a proposed list of candidate dielectrics is presented in Table 2.

In summary, a great deal remains to be learned before the identification of the most attractive alternative gate dielectric. Interesting compositions have been identified, but none has yet been able to overcome all of the associated difficulties. Most certainly, this corresponds to insufficient understanding of the materials in question, the interfaces they form, and the technologies required to produce the sensitive processing equipment needed for accomplishing these aggressive goals. It seems, however, that the issues mentioned above (along with other more subtle factors) are being addressed systematically, and the results of these fundamental investigations represent real progress. But the lessons from BST for DRAMs emphasize that the effort required to achieve this fundamental understanding should not be underestimated.

The outlook

The silicon-based semiconductor industry is facing impending changes if it is to meet the performance projections set out by the ITRS. In the previous section we have described a solution based upon a change of the dielectric material. However, this is clearly not the only possible solution. We postulate that the possible options facing the industry may be summarized as follows: (1) replace the gate dielectric stack with the new materials discussed, but at the same time introducing minimal process changes; (2) replace the gate dielectric stack, and make substantial process changes to reduce process temperatures, but still retain the present CMOS device; (3) change the device geometry, structure and process, but retain the current materials; or (4) introduce entirely new device mechanisms, for example, using alternative physical principles.

Which of these options will emerge as the solution? There is research being undertaken which falls into all four categories. The replacement of the gate dielectric discussed in this article represents the first two of the four options. For the gate dielectric to be replaced by one of these high-permittivity alternatives, much more research is required. As an illustration, it should be noted that not one of these dielectrics have yet been demonstrated to show significantly improved capacitance density over SiO_2 or Si-O-N , while at the same time retaining the necessary transistor properties. The lessons of BST should also be noted — that we require a careful rebuilding of the technology platform on which to base a reliable manufacturing capability. We are concerned that the necessary resources are not being allocated to achieve this objective. This increases the possibility that option 3 will be adopted. The advantage of this third approach is that

Table 2 Candidate dielectrics for gate-stack application by three alternative process routes

Dielectric	Application			Permittivity
	Standard flow	Reverse gate	Epitaxial	
ZrO_2		+		22
HfO_2		+		21
La_2O_3		+		25–30
Al_2O_3		+		11
$\text{ZrO}_2\text{-SiO}_2$	+			12*
$\text{HfO}_2\text{-SiO}_2$	+			11*
$\text{La}_2\text{O}_3\text{-SiO}_2$	+			17†
$\text{Y}_2\text{O}_3\text{-SiO}_2$	+			10†
LaAlO_3			+	25
SrTiO_3			+	200
SrZrO_3			+	25

*These permittivity values belong to the crystalline stoichiometric silicates. Other compositions belonging to this system are also under investigation.

†These values are estimates calculated as a linear mixture of the endmembers for which reference data is known.

it builds upon the existing technology platform, and the approach is fully consistent with the prevailing culture of the industry, namely a process engineering approach, rather than a materials approach to problem solving. As an example of this approach, Hergenrother *et al.* recently fabricated a vertical CMOS FET with a lithography-independent gate length³⁷. Effectively, film thickness (which can easily be controlled on the nano-scale) determines the device dimension, as opposed to lithographic patterns. It is important to note, however, that these structures are non-trivial to manufacture. For example, a level of dopant-diffusion control, currently unavailable, is required for the success of these devices. Thus, even though the design seems very attractive, additional advances in processing are necessary.

What about the possibility that an entirely different type of physical device is developed? Certainly, there is considerable research on alternative device types, and this research is increasing. In addition, end of scaling indicates a possible paradigm shift, an occasion when new 'disruptive technologies' could find their way into the microelectronics mainstream. This is a possibility, although the likelihood is small for any rapid change, owing to the huge investment in the present silicon-based infrastructure.

So what does the future hold? We do not know, but we are excited to be part of the semiconductor industry at a time where significant change is inevitable. □

- Moore, G. E. Progress in digital integrated electronics. *IEEE IEDM Tech. Dig.* 11–13 (1975).
- Moore, G. E. Cramming more components onto integrated circuits. *Electronics* **38**, 114–117 (1965).
- Semiconductor Industry Association. *International Technology Roadmap for Semiconductors 1999* edn <http://www.itrs.net/ntsr/publntrs.nsf>.
- Lo, S.-H., Buchanan, D. A., Taur, Y. & Wang, W. Quantum-mechanical modeling of electron tunneling current from the inversion layer of ultra-thin-oxide nMOSFETs. *IEEE Electron Device Lett.* **18**, 209–211 (1997).
- Summerfelt, S. R. in *Thin Film Ferroelectric Materials and Devices* (ed. Ramesh, R.) 1–42 (Kluwer, Boston, MA, 1997).
- Kotecki, D. E. A review of high dielectric materials for DRAM applications. *Integr. Ferroelec.* **16**, 1–19 (1997).
- Kotecki, D. E. *et al.* (Ba,Sr)TiO₃ dielectrics for future stacked-capacitor DRAM. *IBM J. Res. Develop.* **43**, 339–350 (1999).
- Fazan, P. C. *et al.* Ultrathin oxide nitride dielectrics for rugged stacked DRAM capacitors. *IEEE Electron Device Lett.* **13**, 86–88 (1992).
- Hilton, A. D. & Ricketts, B. W. Dielectric properties of Ba_xSr_{1-x}TiO₃ ceramics. *J. Phys. D* **29**, 1321–1325 (1996).
- Matsubara, S., Sakuma, T., Yamamichi, S., Yamaguchi, H. & Miyasaka, Y. in *Ferroelectric Thin Films* (eds Kingon, A. I. and Myers, E. R.) (*MRS Symp. Proc.* **200**) 243–253 (Materials Research Society, Pittsburgh, PA, 1990).
- Fazan, P. Trends in the development of ULSI DRAM capacitors. *Integr. Ferroelec.* **4**, 247–256 (1994).
- Basceri, C., Streiffer, S. K., Kingon, A. I. & Waser, R. The dielectric response of fiber-textured (Ba, Sr)TiO₃ thin films grown by chemical vapor deposition. *J. Appl. Phys.* **82**, 2497–2504 (1997).
- Streiffer, S. K., Basceri, C., Parker, C. B., Lash, S. E. & Kingon, A. I. Ferroelectricity in thin films: the dielectric response of fiber-textured (Ba_{0.7}Sr_{0.3})TiO₃ thin films grown by chemical vapor deposition. *J. Appl. Phys.* **86**, 4565–4575 (1999).

- Kingon, A. I., Streiffer, S. K., Basceri, C. & Summerfelt, S. R. Application of high-permittivity perovskite thin films to dynamic random access memories. *Mater. Res. Bull.* **21**, 46–52 (1996).
- Baniecki, J. D. *et al.* Dielectric relaxation of Ba_{0.7}Sr_{0.3}TiO₃ thin films from 1 mHz to 20 GHz. *Appl. Phys. Lett.* **72**, 498–500 (1998).
- Horikawa, T., Makita, T. & Mikami, N. Dielectric relaxation of (Ba,Sr)TiO₃ thin films. *Jpn. J. Appl. Phys.* **34**, 5478–5482 (1995).
- Lash, S. E. *Growth and properties of MOCVD (Ba, Sr)TiO₃ thin film capacitors*. Thesis, North Carolina State Univ. (1999).
- Bilodeau, S. M., Carl, R., Van Buskirk, P. & Ward, J. MOCVD of (Ba,Sr)TiO₃ for 1-Gbit DRAMs. *Solid State Technol.* **40**, 235–242 (1997).
- Buskirk, P. C. V. *et al.* Common and unique aspects of perovskite thin film CVD processes. *Integr. Ferroelec.* **21**, 273–289 (1998).
- Basceri, C. *Electrical and dielectric properties of (Ba,Sr)TiO₃ thin film capacitors for ultra-high density dynamic random access memories*. Thesis, North Carolina State Univ. (1997).
- Grossmann, M. *et al.* Resistance degradation behavior of Ba_{0.7}Sr_{0.3}TiO₃ thin films compared to mechanisms found in titanate ceramics and single crystals. *Integr. Ferroelec.* **22**, 603–614 (1998).
- Wu, E. Y., Stathis, J. H. & Han, L.-K. Ultrathin oxide reliability for ULSI applications. *Semiconductor Sci. Technol.* **15**, 425–435 (2000).
- Muller, D. A. *et al.* The electronic structure at the atomic scale of ultrathin gate oxides. *Nature* **399**, 758–762 (1999).
- Hirose, M. *et al.* Fundamental limit of gate oxide thickness scaling in advanced MOSFETs. *Semiconductor Sci. Technol.* **15**, 485–490 (2000).
- Luan, H. F. *et al.* High quality Ta₂O₅ gate dielectrics with T_{meq}>10. *IEDM Tech. Digest Int.* 141–144 (1999).
- Alers, G. B. *et al.* Intermixing at the tantalum oxide/silicon interface in gate dielectric structures. *Appl. Phys. Lett.* **73**, 1517–1519 (1998).
- Campbell, S. A. *et al.* MOSFET transistors fabricated with high permittivity TiO₂ dielectrics. *IEEE Trans. Electron Devices* **44**, 104–109 (1997).
- Wilk, G. D., Wallace, R. M. & Anthony, J. M. Hafnium and zirconium silicates for advanced gate dielectrics. *J. Appl. Phys.* **87**, 484–492 (2000).
- Wilk, G. D. & Wallace, R. M. Electrical properties of hafnium silicate gate dielectrics deposited directly on silicon. *Appl. Phys. Lett.* **74**, 2854–2856 (1999).
- Cheng, Y. C. & Sullivan, E. A. Scattering of charge carriers in silicon. *J. Appl. Phys.* **44**, 923–925 (1973).
- Eisenbeiser, K. *et al.* Field effect transistors with SrTiO₃ gate dielectric on Si. *Appl. Phys. Lett.* **76**, 1324–1326 (2000).
- Hubbard, K. J. & Schlom, D. G. Thermodynamic stability of binary oxides in contact with silicon. *J. Mater. Res.* **11**, 2757–2776 (1996).
- Copel, M., Gribelyuk, M. & Gusev, E. Structure and stability of ultrathin zirconium oxide layers on Si(001). *Appl. Phys. Lett.* **76**, 436–438 (2000).
- Gusev, E. P. *et al.* High-resolution depth profiling in ultrathin AIO₃ films on Si. *Appl. Phys. Lett.* **76**, 176–178 (2000).
- Shannon, R. D. Dielectric polarizabilities of ions in oxides and fluorides. *J. Appl. Phys.* **73**, 348–366 (1993).
- Schlom, D. G. High-K candidates for use as the gate dielectric in silicon MOSFETs. *Appl. Phys. A* (in the press).
- Hergenrother, J. M. *et al.* The vertical replacement-gate (VRG) MOSFET: a 50-nm vertical MOSFET with lithographically-independent gate length. *IEDM Tech. Digest* 75–78 (1999).

Acknowledgements

Research by the authors on the topic of dielectrics for DRAMs and gate dielectrics is funded by SRC and Sematech. Research results contributed by students D. Wicaksana and J. Parrette are gratefully acknowledged. Some data presented are from research undertaken by the US Ultradense Capacitor Materials Processing Partnership. Thanks to C. Parker and C. Osburn for assistance with the figures, and to R. Amos (IBM) for a critical reading of the manuscript. S.K.S. acknowledges support by the US Department of Energy, Office of Science.

Amplifying quantum signals with the single-electron transistor

Michel H. Devoret*† & Robert J. Schoelkopf*

*Department of Applied Physics, Yale University, New Haven, Connecticut 06520, USA

†Service de Physique de l'Etat Condensé, CEA-Saclay, F-91191, France

Transistors have continuously reduced in size and increased in switching speed since their invention in 1947. The exponential pace of transistor evolution has led to a revolution in information acquisition, processing and communication technologies. And reigning over most digital applications is a single device structure — the field-effect transistor (FET). But as device dimensions approach the nanometre scale, quantum effects become increasingly important for device operation, and conceptually new transistor structures may need to be adopted. A notable example of such a structure is the single-electron transistor, or SET^{1–4}. Although it is unlikely that SETs will replace FETs in conventional electronics, they should prove useful in ultra-low-noise analog applications. Moreover, because it is not affected by the same technological limitations as the FET, the SET can approach closely the quantum limit of sensitivity. It might also be a useful read-out device for a solid-state quantum computer.

It is now possible to put a billion transistors on a single chip operating with a clock period of a billionth of a second. Most probably, the trend in reducing dimensions and times will continue in the next decade. But as the number and density of gates and memory elements increase, the energy of signals also has to be reduced to keep the power dissipation sufficiently low.

Surprisingly, even though the size of a typical transistor in a microcomputer chip is now just a few hundred nanometers, its functioning remains essentially classical: quantum mechanics only enters in the explanation of the values of the physical parameters of materials, like the band-gap of a semiconductor. Otherwise, the discreteness of matter and the wave-like properties of electrons can be largely ignored in the understanding of the behaviour of electrical signals in today's integrated circuits.

But as devices get smaller, faster and more densely packed, quantum effects will have increasingly to be taken into account. Even well before we reach the ultimate limit where transistors are reduced to the size of an atom or a molecule, we encounter four limits. Quantum phenomena become significant when (1) signal energy, (2) signal charge, (3) device dimension, and (4) device size tolerance approach, respectively, the energy of one photon, the charge of one electron, the electron wavelength, and the size of one atom.

Much research has been devoted to assess if quantum effects arising from these conditions will force the adoption of new physical principles or if they can simply be tamed by better control of the chip structure at the atomic level. To our knowledge, there is no general consensus on the answer to this question.

Another research direction has been to exploit quantum effects arising in devices of nanometer scale to implement a function that cannot be performed by present devices. In some applications, which operate at limits (1) or (2) or both, it is not only inevitable but also desirable. In astronomy, for instance, it is important to extract as much information as possible from a single photon⁵. Recent advances in quantum information theory⁶ indicate that scalable switching elements that behave fully as quantum systems would not

simply make calculations with minimal energy, they could in addition perform tasks that would be impossible with conventional computers. In a quantum computer, usual bits are replaced by quantum bits or 'qubits' which can be 'entangled' with each other, thereby carrying a new type of information that is useful in solving highly parallel tasks. New types of devices are needed to read-out such qubits, that is, to amplify their associated single-quantum signals.

In the realm of atomic physics and quantum optics, the detection of individual microscopic particles travelling in vacuum, such as photons and electrons, is now performed routinely with almost unity efficiency by instruments derived from the photomultiplier. However, the measurement of electrical signals resulting from the motion of a single electron in a circuit involves an amplifier having not only a good energy sensitivity, but also electrical characteristics that are adapted to this circuit.

A particularly simple and spectacular example of such a device is the single-electron transistor (SET)^{1–3}, which exploits the quantum phenomenon of tunnelling to control and measure the movement of single electrons inside a solid-state circuit. SETs are extremely precise solid-state electrometers^{4,7}, already out-performing state-of-the-art conventional transistors⁸ by three orders of magnitude. Their charge sensitivity has been shown to be as low as a few $10^{-5} e/\sqrt{\text{Hz}}$, which means that a charge variation of $10^{-5} e$ can be detected in a measurement time of 1 s (the precision improves as the square root of the measurement time). This result is only an order of magnitude away from the theoretical limit of $10^{-6} e/\sqrt{\text{Hz}}$. SETs have applications in metrology⁹ and single-photon detection^{10,11}. Furthermore, it has been realized in the past few years that SETs can perform a measurement on a single quantum two-state system (qubit)^{12,13}, perturbing its quantum evolution in a minimal way. That is, the SET is a charge amplifier operating in the vicinity of the quantum limit. It would be a practical read-out device for several solid-state implementations of qubits¹⁴. In this article we review these latest developments.

Conventional transistors

Before discussing the performance of SETs, it is useful to recall the operating principle of the most common transistor,

the metal-oxide semiconductor field-effect transistor (MOSFET). Figure 1 depicts schematically the layout of the device and its operating principle (we restrict ourselves here to the nMOSFET, in which the majority charge carriers are electrons). Two conducting electrodes, called the source and drain, are connected by a channel made of a material in which the number of conduction electrons can be varied, in practice a semiconductor (Fig. 1a). A voltage is applied to the 'gate', a third conducting electrode that is separated from the channel by a thin insulating layer. When the gate–source voltage is zero, there are no conduction electrons in the channel as the effective potential they would experience there is larger than in the leads (Fig. 1b). The channel is therefore in an insulating state. But when the gate voltage is increased with respect to the source, the potential experienced by the electrons in the channel decreases and they populate the channel just under the gate (Fig. 1c). The channel becomes conducting. The larger the gate voltage, the larger will be the channel electron population that can participate in the current. Eventually, all the electron states in the energy window set by the source–drain voltage can propagate through the channel and the current no longer depends on the gate voltage. This saturation regime is depicted in Fig. 1d.

This field effect provides an amplification mechanism since an increase in gate voltage, bringing a modest current to the gate electrode, can switch on a larger current through the channel (Fig. 2). The source–drain current is determined by the conductance of the channel, which in turn depends on two factors: the density of its conduction electrons and their mobility. The electron density is controlled directly by the gate voltage. The electron mobility is set by the collisions of electrons with static irregularities of the crystal as well as with its dynamic deformations due to thermal agitation. When thermal agitation is the predominant factor, electron mobility is, to a large extent, independent of the gate voltage. However, at low temperatures, mobility can also increase when the density increases, reinforcing the influence of the gate voltage on the source–drain current.

Note that so far we have made no reference to the wave-like properties of electrons nor to the fact that the channel is made from individual atoms. The only quantum property that has had a role in our explanation is the Pauli principle, which dictates that each possible state for an electron in the channel can be occupied at most by only one electron. This means that only a certain number of electrons can accumulate in the channel, setting a limit on the current flow.

However, the quantum properties of electrons and atoms will be increasingly important as FETs are made smaller. For example, the wave nature of electrons will influence the way they travel through the channel. When the transverse dimension of the channel becomes comparable to the wavelength of electrons (around 100 nm), electron propagation becomes more sensitive to the atomic disorder in the device, which is inherent in the present fabrication process. The disorder makes the channel remain insulating even when the density of electrons is increased. Effects of this kind, which result from reaching the limit to device dimension (limit (3) above), pose a major problem if the reduction of size is not accompanied by an improvement in the atomic structure of the fabricated devices.

If, however, the atomic structure of the FET could be made defect-free, a recent analysis¹⁵ shows it would continue to function in the regime where the electron propagation in the channel is wave-like. It would thus 'break' limit (3) and could be further scaled down until a channel length of about 8 nm is reached, a stage at which limit (4) seems to severely affect the performance of the FET.

In an ideal situation, we might want electrons to be scattered only by the gate-dependent potential, with other scattering mechanisms always being detrimental to amplification. But the confinement of electrons in the channel by tunnel barriers, in the source–drain direction, can also lead to a new kind of amplification principle. This is the basis of the SET. As we shall see, this new principle circumvents the problem of the FET that the gate capacitance, which is an important factor in determining charge sensitivity, is tied to the size of the

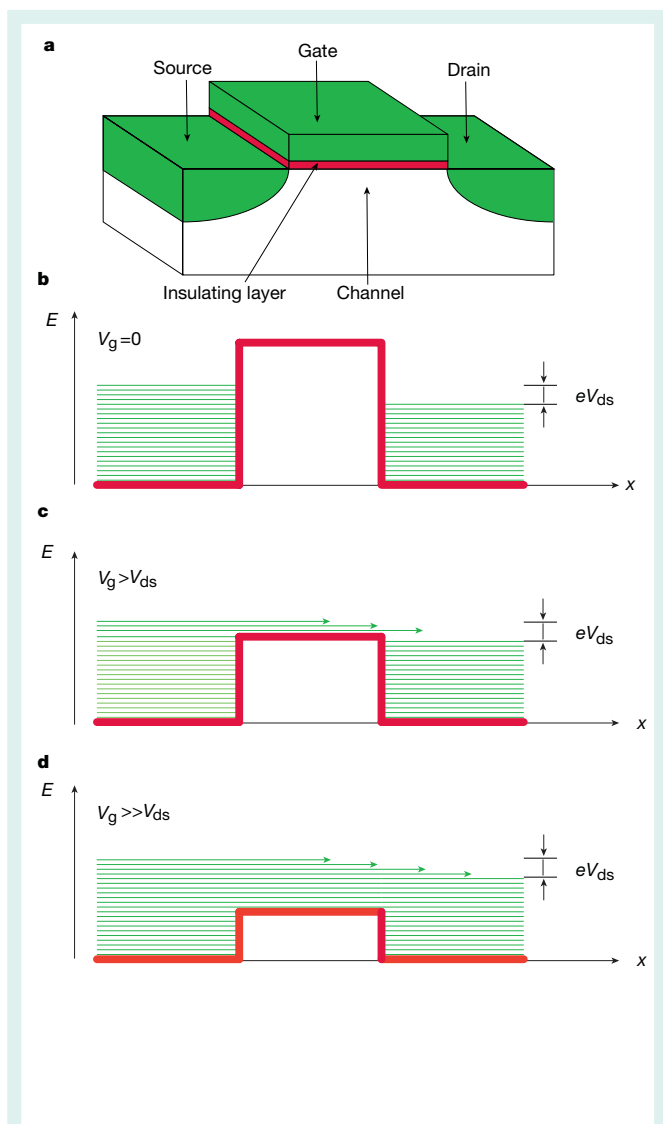


Figure 1 Principle of a metal-oxide semiconductor field-effect transistor (MOSFET).

a. The device consists of two conducting electrodes (source and drain) connected by a semiconducting channel. The channel is influenced electrostatically by the gate, a conducting electrode separated from the channel by thin insulating layer. **b.** When the voltage applied between the gate and the source is zero, the Fermi energy of the source and drain lies in the gap of the semiconductor. Here, we sketch the potential (red curve) seen by conduction electrons when they travel from source to drain along a line in the channel just under the gate. There are no filled electron states (green lines) in the channel, which as a result remains insulating. **c.** When the gate voltage is increased, the potential seen by conduction electrons is lowered. There are now filled states in the channel at the Fermi energy of the source and drain. The device conducts. **d.** When the gate voltage is increased further, the current finally saturates when all the states in the bias window are filled.

channel. Until the technology for reliably fabricating FETs with channels of nanometre-scale dimensions has been developed, the SET is the best device in terms of charge sensitivity.

Operating principle of SETs

Unlike the FET, whose principle does not require the motion of electrons to be quantum-mechanical, the SET is based on an intrinsically quantum phenomenon: the tunnel effect through a metal–insulator–metal junction. When two metallic electrodes are separated by an insulating barrier whose thickness is only ~ 1 nm, electrons at the Fermi energy can traverse the insulator even though their energy is too

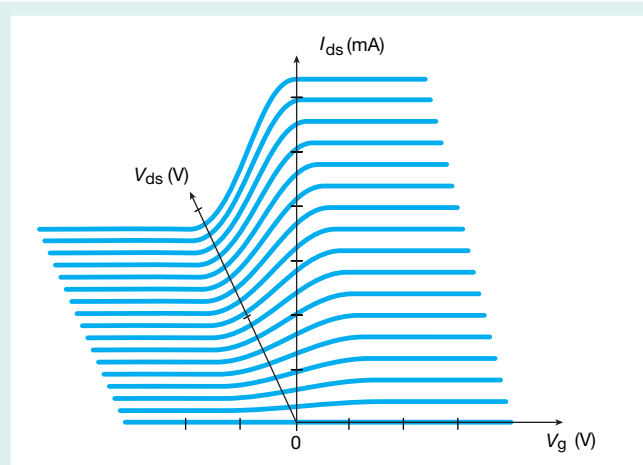


Figure 2 Variation of the source–drain current in a MOSFET as a function of the gate voltage. When the gate voltage is increased from zero, the source–drain current is turned on. This device can be used both in digital electronics and as an amplifier for analog signals.

low to overcome, in a classical motion, the large potential barrier of the insulating region. The tunnel effect manifests itself by a finite resistance R_T of the insulating barrier. This resistance depends both on the transmission coefficient $\mathcal{T}$ of the barrier to electron waves (which is an exponentially decreasing function of its thickness) and on the number M of independent electron wave modes impinging on the barrier (this number is equal to the area of the junction divided by the square of the electron wavelength). The SET uses a key property of the tunnel effect in a many-electron system: for barriers such that $\mathcal{T}M \ll 1$, the charge Q transferred through the barrier becomes quantized with $Q = Ne$, where N is an integer¹⁶. In other words, for N not to be subject to quantum fluctuations, the resistance of the junction must be large compared with the resistance quantum $R_T \gg h/e^2 = R_K = 25.8 \text{ k}\Omega$ (refs 17,18).

The SET consists of two such tunnel junctions placed in series (Fig. 3a,b). An ‘island’ is thus formed between the two junctions. A gate electrode is coupled electrostatically to the island. The SET can thus be described as a FET in which the semiconducting channel has been replaced by a metallic island sandwiched between two tunnel barriers. The island has a total capacitance C_S , which is the sum of the gate and junction capacitances $C_S = C_g + C_{J1} + C_{J2}$.

If the dimensions of the island are sufficiently small, the charging energy $E_C = e^2/(2C_S)$ of one extra electron in the island will become larger than the energy of thermal fluctuations: $E_C \gg k_B T$, where k_B is the Boltzmann constant and T is temperature. In practice, for devices fabricated by standard electron-beam lithography, C_S is of the order of a femtofarad and the charging energy is of order 1 K, necessitating temperatures below 300 mK to satisfy the above charging energy criterion. Over the past few years, however, experiments have shown that with advanced fabrication methods, room temperature operation is possible^{19–21}.

Because electrons interact strongly via the Coulomb interaction when they pass through the island, the analysis of the SET differs fundamentally from that of the FET. In the FET, electrons go from source to drain independently, and in such numbers that one can consider that the potential seen by one is an average which does not depend on the configuration adopted by all of the others. Electrical transport results from a simple addition of the motion of each electron. In the SET, by contrast, transport results from transitions between collective charge states of the system. These charge states are described by the two numbers N_1 and N_2 of electrons having traversed the junctions (Fig. 3b).

The behaviour of the device is governed by the global electrostatic

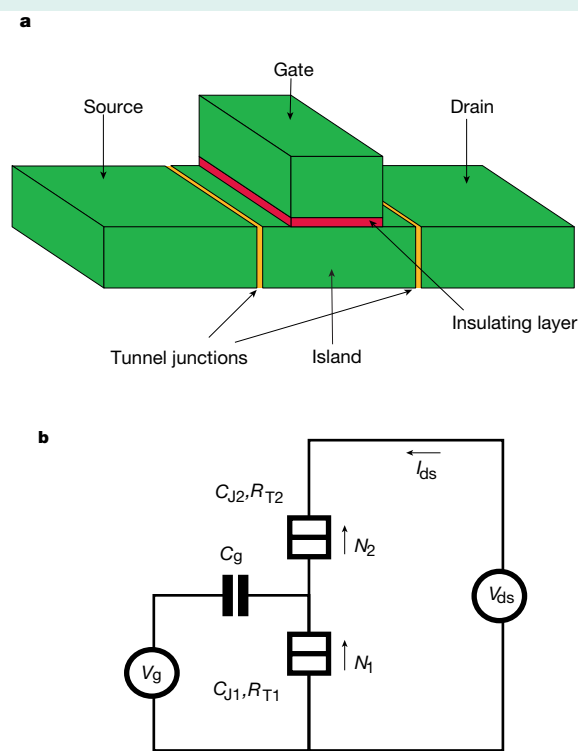


Figure 3 The single-electron tunnelling transistor (SET). **a**, Simplified three-dimensional structure of the SET. The channel of the FET is replaced here by a sandwich consisting of a nanoscale metal electrode (island), which is connected to the drain and the source by tunnel junctions. As in the FET, a gate electrode influences the island electrostatically. **b**, Circuit diagram of the SET. The square box symbol represents a tunnel junction, and integers N_1 and N_2 denote the numbers of electrons having tunneled through the two junctions. Each junction is characterized by its capacitance and its tunnel resistance.

energy¹⁶ $E_{cl} = E_C[N_2 - N_1 - (C_g V_g/e) - (C_2 V_{ds}/e) + q_0]^2 - eN_2 V_{ds}$, which includes the energy stored in the junction and gate capacitances, as well as the work done by the voltage sources. Here, V_g and V_{ds} are the voltages applied between gate and source, and drain and source, respectively. The so-called offset gate charge q_0 is a phenomenological quantity describing the fact that electric fields in the capacitances of the system are non-zero even when the island is neutral and when no voltage is applied. It takes a randomly different, non-integer value for each device and cool-down. It also fluctuates slowly in time with a $1/f$ spectral density²². We will discuss its effect in more detail below. But as far as the amplification mechanism of the SET is concerned, we can treat it as a constant.

According to the so-called ‘global rules’, also known as ‘orthodox theory’, tunnel events will take place independently on each junction at a rate governed by the global energy, provided that the junction resistances satisfy $R_{T1}, R_{T2} \gg R_K$ and that the voltage sources V_g and V_{ds} have negligible internal impedance, on the scale of the resistance quantum, around the Coulomb frequency E_C/h (ref. 17).

In this regime, each tunnel event creates one electron–hole pair, the electron and the hole being on opposite sides of the junction. The succession of tunnel events constitutes a Poisson process. More specifically, a tunnel event will take place on junction i with a rate given by $\Gamma_i = [1/(R_{Ti} e^2)] [\Delta E_i / (1 - \exp(-\Delta E_i/k_B T))]$ where $\Delta E_i = E_{cl}\{N_i^{\text{before}}, N_j\} - E_{cl}\{N_i^{\text{after}}, N_j\}$.

At zero temperature, tunnel events take place only if they are energetically allowed, that is, $\Delta E_i > 0$. For a drain–source voltage below the Coulomb gap voltage e/C_S , the current therefore depends critically on the value of the gate voltage. If the gate voltage is such that

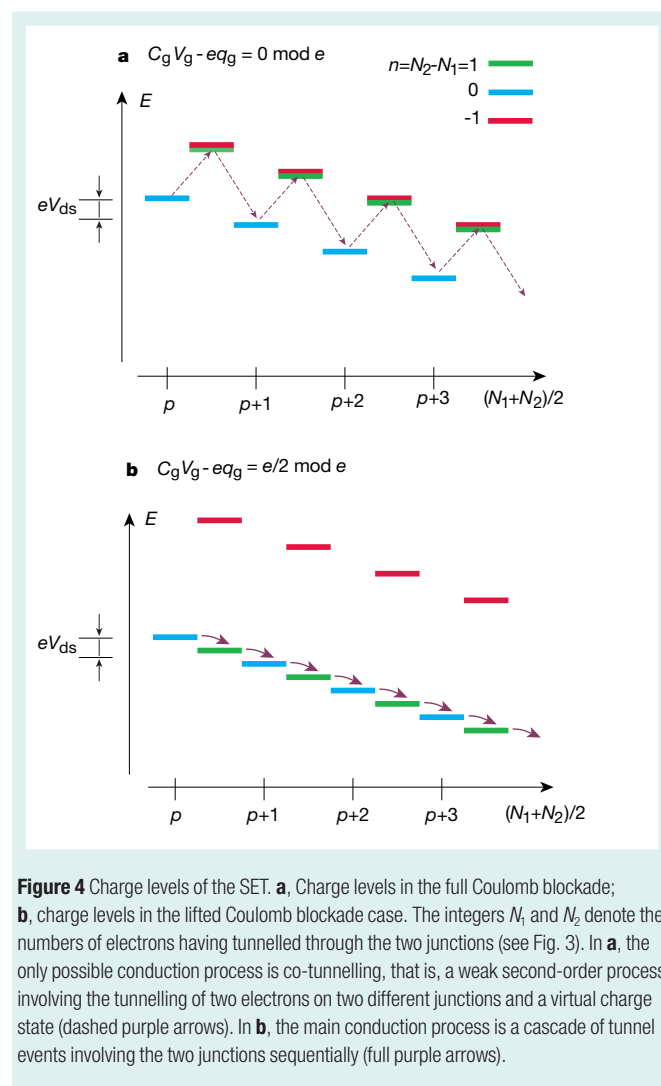


Figure 4 Charge levels of the SET. **a**, Charge levels in the full Coulomb blockade; **b**, charge levels in the lifted Coulomb blockade case. The integers N_1 and N_2 denote the numbers of electrons having tunneled through the two junctions (see Fig. 3). In **a**, the only possible conduction process is co-tunnelling, that is, a weak second-order process involving the tunnelling of two electrons on two different junctions and a virtual charge state (dashed purple arrows). In **b**, the main conduction process is a cascade of tunnel events involving the two junctions sequentially (full purple arrows).

$C_g V_g = q_0 e \bmod e$, the island has a well defined number of electrons. Tunnel events cannot take place on the junctions because the global energy would increase (Fig. 4a) and the current is zero. This is the phenomenon of Coulomb blockade. But if the gate voltage is such that there is a charge equivalent to one half electron on the gate capacitance, $C_g V_g = (q_0 + \frac{1}{2})e \bmod e$, tunnel events are energetically allowed (Fig. 4b). A cascade of transitions between charge states occurs, and current flows between source and drain.

Note that in practice the current cannot be turned off completely in the Coulomb-blockaded state. Higher-order processes in which several tunnel events occur simultaneously (co-tunnelling processes), and whose relative importance scale as powers of R_K/R_T , will induce a small leakage current not described by the orthodox theory. However, as long as $R_K \gg R_T$, the SET behaves as a charge amplifier as the presence or absence of a fraction of an electron on the gate capacitance can control an easily measurable current (the order of magnitude is around 10^9 e/s in practical cases) (Fig. 5). This gain is not the only factor determining the sensitivity of the device. One must take into account the shot noise in the source–drain current, which is attributable to the quantum randomness of the time intervals between tunnel events (we work in a regime where thermal fluctuations can be neglected). This quantum randomness, corresponding to electrons having to ‘choose’ between the two sides of the barrier, is the process that ultimately limits the sensitivity of the device. As we will show below, the noise characteristics of the SET can be calculated exactly in a simple regime, which is rich enough to yield semi-quantitative understanding of the quantum limit of sensitivity.

Returning to the offset gate charge, it is believed that its value is determined in part by differences between the work functions of the metal of the island and that of the other electrodes. Even minute variations in the work functions are sufficient to cause q_0 to fluctuate by numbers much greater than unity, which is the case observed in practice. Another potential source of fluctuation in offset charge is charge motion in the substrate or even in the tunnel barrier oxide. The absence of control over offset charges severely hinders any application of SETs to digital electronics, where the gate thresholds must be rigorously fixed and uniform. However, SETs can still be used as sensitive amplifiers in the audio-frequency (a.f.) or radio-frequency (r.f.) domains, as a simple additional feedback circuit can compensate for the fluctuations in offset charge that occur mainly at lower frequencies.

Noise characteristics of an amplifier

Before we examine how quantum shot noise affects the ultimate performances of SETs, it is useful to discuss the noise properties of a general amplifier.

A linear amplifier can be described phenomenologically as a ‘black box’ with two input leads and two output leads. We can represent the inside of the black box by effective elementary components which accurately describe how it appears to the outside circuitry. If we limit ourselves to the simpler case where the amplifier has no feedback (that is, input current is independent of output voltage), we arrive at the schematics of Fig. 6 (feedback introduces complications that are not crucial in our discussion). Three elements describe the transformation of the signal by the amplifier: an ideal voltage amplifier with infinite input impedance, zero output impedance and a voltage gain $G(\omega)$, and the input and output impedances, $Z_{in}(\omega)$ and $Z_{out}(\omega)$. In these parameters, the argument ω denotes the signal angular frequency.

In addition, the random fluctuations due to the amplifier are described by two noise sources with very different roles. The voltage noise source V_N describes the output noise, that is, the noise added by the amplifier to the output signal, referred to the input. The current noise I_N , on the other hand, describes the back-action of the amplifier on the circuit at the input. The voltage and current noise are assumed to be gaussian, and are characterized by the spectral densities $S_V(\omega)$ and $S_I(\omega)$, which are the Fourier transforms of the autocorrelation functions of V_N and I_N , respectively. We neglect here the correlations $S_{VI}(\omega)$, which are not essential for our discussion. These spectral densities, together with the input impedance, determine the ultimate resolving power of the amplifier for small signals.

It is useful, in the discussions of this resolving power, to introduce several combinations of these quantities. The first one is the charge sensitivity $\delta Q(\omega) = \sqrt{S_V(\omega)/|\omega Z_{in}(\omega)|}$, which we have already discussed. The second quantity is the energy sensitivity $\epsilon(\omega) = \delta Q^2(\omega)/2C_{in}$ where $C_{in} = \text{Re}\{1/[i\omega Z_{in}(\omega)]\}$. Assuming that the amplifier is driven by a voltage source with strictly zero internal impedance (and hence always in the classical regime, as the zero-point voltage fluctuations are proportional to the real part of the impedance), this quantity tells us the smallest amount of energy $\delta E = \epsilon/\tau$ fed by the source into the input circuit of the amplifier which will give an output signal that is distinguishable from zero when we accumulate it during a time τ . Even though ϵ has the dimension of an action and is conveniently measured in units of $\hbar$, quantum mechanics does not impose any restriction on the ratio $\epsilon/\hbar$, which can in principle tend to zero²³. Nevertheless, in systems studied so far, a value of $\epsilon/\hbar$ of order unity usually means that the contribution of thermal fluctuations to the noise of the amplifier have been suppressed down below those of quantum fluctuations.

Although the energy sensitivity seems initially to be a very powerful concept, it neglects the back-action of the amplifier and is insufficient to determine how well the amplifier performs when we use it to measure a quantum signal coming from a system with a finite source impedance Z_s . The current noise, which induces back-action voltage fluctuations across the source impedance, and hence on the

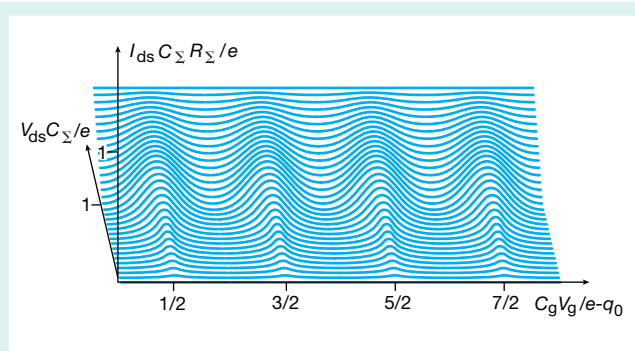


Figure 5 Variation at $T = 0$ of the source–drain current in a SET as a function of the voltage between drain and source and the voltage between gate and source. We have assumed that the two junctions have identical parameters. The sum of the junction tunnel resistances R_Σ is twice the resistance quantum h/e^2 and a small amount of co-tunnelling rounds the peaks associated with the lifting of the Coulomb blockade (gate charge corresponding to half-integers multiplied by the charge quantum). The sensitivity of the SET is based on the rapid variation of the source–drain current as the gate charge varies by a fraction of one electron.

signal going into the amplifier, must be considered together with the voltage noise.

We therefore introduce the noise impedance $Z_N = \sqrt{S_V/S_I}$ and the noise energy $E_N = \sqrt{S_V S_I}$. They have the following meaning: supposing that $1/Z_{in}(\omega) = 0$, the amplifier will add a minimal amount of noise to the signal coming from the source when $Z_S = Z_N$ (noise matching). This minimal noise has a power per unit bandwidth given by E_N (this quantity has the dimension of an energy, hence the name noise energy). Quantum mechanics places a strict restriction on the noise energy. No amplifier can have an E_N smaller than $\hbar\omega/2$, half a photon energy at the signal frequency ω (refs 23–25). This fundamental limitation is a form of Heisenberg's uncertainty principle (see Table 1). It is worth mentioning that in the classical regime, that is, the regime where the contribution of quantum fluctuations to the noise of the amplifier is negligible, the noise temperature $T_N = E_N/k_B$ is often used in place of the noise energy. The meaning of T_N corresponds to a well-defined procedure: imagine one connects a resistor with value Z_N at the input of the amplifier. The temperature to which one must heat the resistor to double the noise measured at the output of the amplifier is precisely T_N .

If we now apply these concepts to FET amplifiers, we find that the best performance in the r.f. domain 0.1–10 GHz is obtained with heterostructure FETs cooled to 4 K (ref. 26). At 500 MHz, their noise temperature is around 2 K (the corresponding thermal energy is equivalent to the energy of 40 photons at 1 GHz), their noise impedance is about 50 Ω and their energy sensitivity is of the order $10^2 \hbar$. They are thus far from the quantum limit (at higher frequencies the minimum noise energy improves and can be equivalent to only ten photons). As a charge-sensing device, their best performance is around $10^{-2} e/\sqrt{\text{Hz}}$ (ref. 8).

Although theoretically the heterostructure FET could reach the quantum limit of sensitivity in the form of a ballistic, two-dimensional electron gas quantum-point contact²⁷, it is difficult in

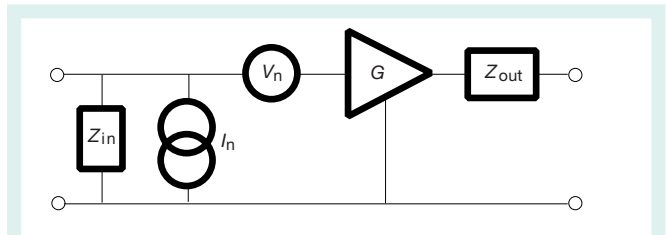


Figure 6 Effective circuit elements describing the properties of a linear voltage amplifier with no feedback. The triangle represents an ideal noiseless voltage amplifier with infinite input impedance and zero output impedance. Its gain is denoted by G . The boxes marked Z_{in} and Z_{out} correspond to the input and output impedances, respectively. The noise sources V_n and I_n represent the voltage noise (noise added by the amplifier at the output, but referred to the input) and the current noise (back-action noise of the amplifier sent to the circuit feeding the input). These elements are eventually dependent on the signal frequency ω

this device to achieve effective electrostatic coupling of the gate electrode to the channel without affecting its mobility and inducing parasitic input capacitance. In the SET, by contrast, there is more flexibility in the design of the gate capacitance. The fabrication of the latter is to a large extent disconnected from that of the tunnel barriers, whose quality is analogous to the mobility of the channel in the FET. We will now examine how the operating principle of the SET brings us close to the quantum limit.

Ultimate sensitivity of SETs

In the framework of the orthodox theory, the tunnel events constitute a generalized Poisson process with correlations. It is possible to calculate analytically the noise characteristics of the SET at temperatures such that $k_B T \ll eV_{ds}$ (refs 7, 28). The following expressions are obtained for the voltage and current noise, in the simple case where the two junctions have identical parameters:

$$S_V(\omega) = \frac{(1 - \alpha^2)(1 + \alpha^2)}{8\alpha^2} eV_{ds} R_\Sigma \left(\frac{C_\Sigma}{C_g} \right)^2$$

$$S_I(\omega) = \frac{(1 - \alpha^2)}{4} \frac{eV_{ds}}{R_\Sigma} \left(\frac{C_g}{C_\Sigma} \right)^2 \left(\frac{e\omega R_\Sigma}{V_{ds}} \right)^2$$

In what follows, we neglect the influence of the correlation

$$S_{VI}(\omega) = \frac{(1 - \alpha^2)}{8} eV_{ds} \left(\frac{i\omega e R_\Sigma}{V_{ds}} \right)$$

between voltage and current noise, whose effect would be to reduce the noise energy if an appropriate complex source impedance could be presented to the SET.

In these noise expressions the Coulomb blockade parameter $\alpha = (2C_g V_g - e)/C_\Sigma V_{ds}$, which is limited here to the range $0 < \alpha < 1 - \text{Max}(R_K/\pi R_\Sigma, eV_{ds}/k_B T)$, fixes the relative values of the gate and

Table 1 Constraints on sensitivity and back-action imposed by quantum mechanics

System	Sensitivity	Back-action	Limiting relation	Limited quantity
Heisenberg microscope	Δx	Δp	$\Delta x \Delta p \geq \hbar/2$	Action
Electronic amplifier	S_V	S_I	$(S_V S_I)^{1/2} \geq \hbar\omega/2$	Noise energy
Qubit read-out	T_m	Γ_ϕ	$T_m \Gamma_\phi \geq 1/2$	Information

Table 1 shows different forms of Heisenberg's uncertainty relation linking the sensitivity of a measurement of a given physical quantity and the simultaneous back-action on the conjugate quantity. The precision with which the position of an object would be measured with a photon is related to the momentum transferred to this object by the radiation pressure of the photon (row 1). Similarly, the voltage sensitivity with which an amplifier measures the circuit at its input is related to the back-action current noise that this amplifier produces in the circuit (row 2). Finally, the time needed to acquire the value of a qubit is related to the rate of dephasing of the qubit induced by the back-action of the read-out (row 3). Quantities characterizing sensitivity and back-action, respectively, are given in columns 2 and 4.

drain–source voltage. For $1 - R_K/(\pi R_\Sigma) < \alpha$, co-tunnelling processes dominate over the single tunnelling processes we consider here, and for $1 - eV_{ds}/k_B T < \alpha$ the influence of thermal fluctuations would have to be taken into account in the above expressions. The resistance $R_\Sigma = R_{T1} + R_{T2}$ is the sum of the two junction resistances. The above expressions are valid for source–drain voltages below the Coulomb gap $V_{ds} C_\Sigma / e < 1$ and at frequencies ω that are low on the tunnel rate scale $(eR_\Sigma/V_{ds})^{-1}$. We have taken the mean offset charge to be zero because it appears only as a shift in V_g , but the typical value of the expected $1/f$ offset charge fluctuations make our expressions valid only above the crossover to the intrinsic device shot noise at a few 10^5 Hz. Note that at this level of approximation the input impedance is simply a capacitance resulting from the series combination of the gate and the sum $C_J = C_{J1} + C_{J2}$ of the two junction capacitances: $Z_{in} = C_\Sigma / [i\omega C_\Sigma (C_{J1} + C_{J2})]$. A dissipative part in the input impedance appears only at higher orders in the dimensionless signal frequency $(e\omega R_\Sigma/V_{ds})^{-1}$.

It is straightforward to go from these expressions to the charge sensitivity, energy sensitivity, noise energy and noise impedance:

$$\delta Q(\omega) = \frac{C_J}{\alpha} \sqrt{\frac{(1 - \alpha^4)}{8} \left(\frac{R_\Sigma}{R_K} \right) \left(\frac{\hbar V_{ds}}{e} \right)}$$

$$\epsilon(\omega) = \frac{\pi \hbar (1 - \alpha^4)}{8 \alpha^2} \left(\frac{R_\Sigma}{R_K} \right) \left(\frac{V_{ds}}{e C_\Sigma} \right) \left(\frac{C_J}{C_g} \right)$$

$$E_N(\omega) = \frac{\pi (1 - \alpha^2)}{2 \alpha} \sqrt{\frac{(1 + \alpha^2)}{2}} \frac{R_\Sigma}{R_K} \hbar \omega$$

$$Z_N(\omega) = \sqrt{\frac{1 + \alpha^2}{8 \alpha^2}} \left(\frac{C_\Sigma}{C_g} \right)^2 \frac{V_{ds}}{e \omega}$$

We can find a conservative upper bound for the optimal value of the quantities that characterize the sensitivity of the SET by minimizing the above expressions over the range of parameter values that correspond to our hypotheses. We therefore take $\alpha = 1 - 2R_K/3R_\Sigma$ and $R_\Sigma = 2R_K$, values at which co-tunnelling remains marginal. We also take the source–drain voltage $V_{ds} = e/(2C_\Sigma)$ to be able to neglect thermal fluctuations in practical situations. In addition, we take $C_J = 1$ fF as a compromise between reaching attainable temperatures, keeping the heating of the island by the drain–source current at a reasonable level and achieving an acceptable output signal (in contrast with the previous parameter choices, this last value is dictated essentially by the current technology issues and not by the validity of our expressions). We arrive at the optimal values

$$\delta Q_{opt} \approx 1.7 \times 10^{-6} e / \sqrt{\text{Hz}}$$

$$\epsilon_{opt} \approx 0.7 \hbar (C_J/C_g)$$

$$E_{N,opt} \approx 2.2 \hbar \omega$$

$$[G \times E_N]_{opt} = 0.14 e V_{ds} \approx 75 \text{ mK} \times k_B$$

We have left the internal coupling ratio (C_J/C_g) of the SET in the right-hand side of the expression for ϵ_{opt} as this parameter can easily be modified by changing the lithography of the device. This ratio determines how much of the energy fed by the gate voltage into the SET is used to charge the island, which is the active element, rather than the gate capacitance.

The value for $E_{N,opt}$ shows that the SET operates in the vicinity of the quantum limit. The product $[G \times E_N]_{opt}$ gives the SET output noise which would ideally be higher than the added noise of the following amplifier, that is, a cryogenic heterostructure FET. Today, this FET is the limiting factor in the system noise, and the best experimental upper bound so far for the energy sensitivity is $\epsilon \approx 40 \hbar (C_J/C_g)$

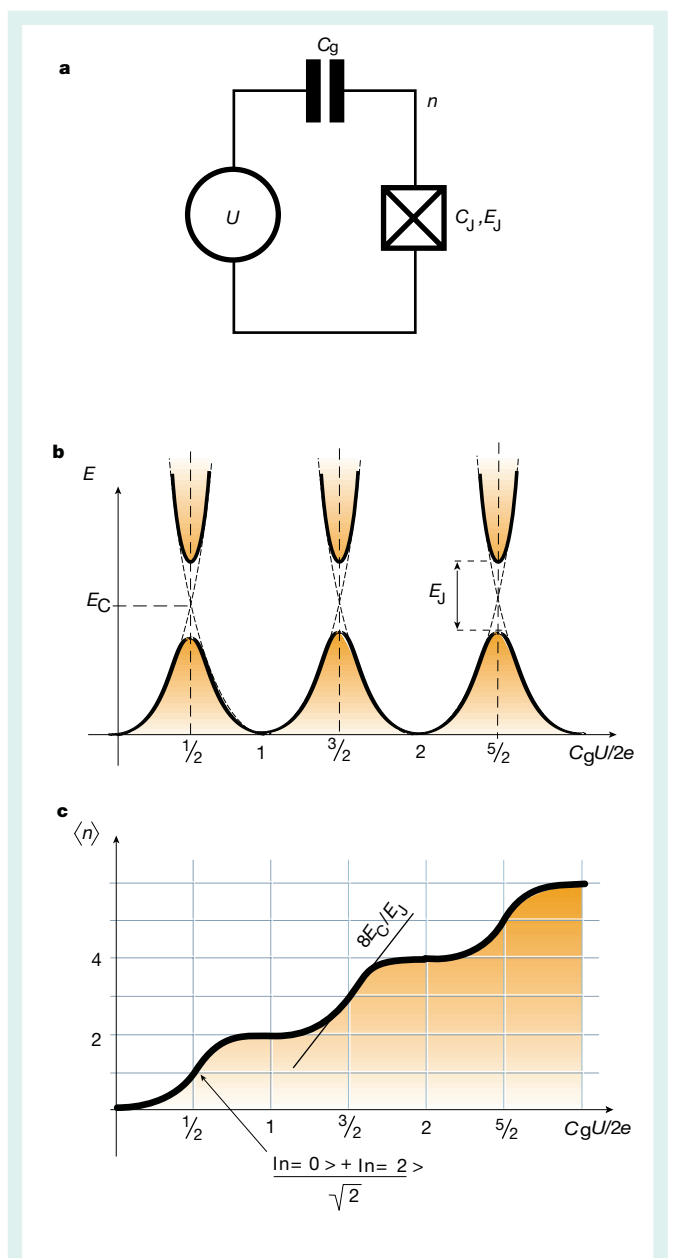


Figure 7 The single-Cooper-pair box. **a**, Circuit diagram of the single-Cooper-pair box. This simple system consists of one superconducting tunnel junction (box with cross) in series with a voltage source U and a gate capacitance C_g . When the charging energy of the island is less than the superconducting gap, all electrons are paired and the number n of excess electrons is even. Cooper pairs can tunnel reversibly through the tunnel junction. Charge states differing by one Cooper pair are coupled by the Josephson energy E_J . **b**, Energy of the two lower quantum states of the Cooper-pair box (full line) as a function of U . The dashed lines represent the electrostatic energy of the box including the work done by the voltage source. At the avoided crossings, the two charge states are mixed and constitute a solid-state implementation of a qubit. **c**, Quantum mechanical average of the number of electrons in the ground state of the box, as a function of U .

in the r.f. domain⁴. The improvement of the combination of the SET with a FET, or with a superconducting quantum interference device (SQUID) amplifier (see below), as well as the verification of the above predictions for the noise, is a topic for future research.

Other quantum-limited, solid-state r.f. amplifiers

According to the above analysis, the SET approaches but does not quite reach the quantum limit. The back-action noise due to the

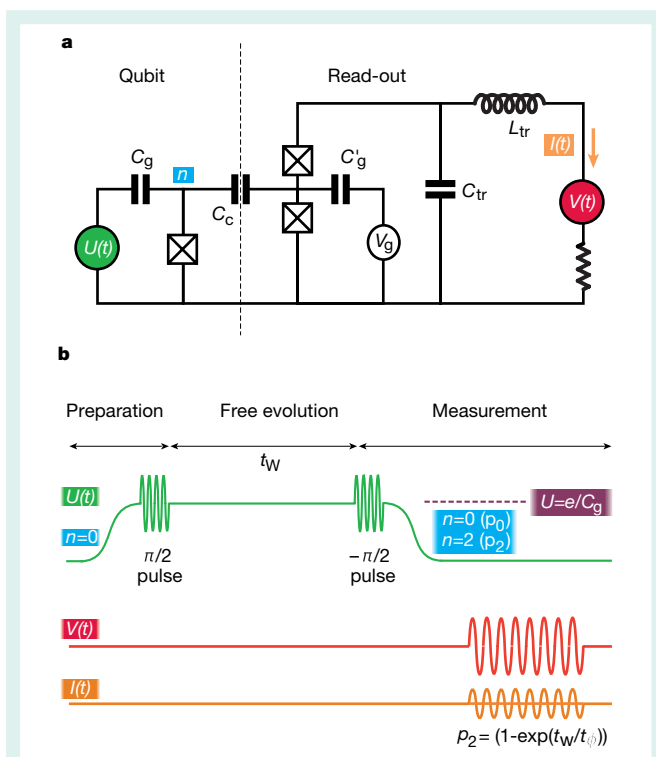


Figure 8 Measuring the state of a quantum system. **a**, Schematic circuit showing how a qubit implemented by a Cooper-pair box could be read-out by an r.f.-SET. **b**, Time evolution of signals in the circuit of **a**. While the read-out r.f.-SET is turned off, a first pulse prepares the qubit in a quantum superposition of ground and excited state at the crossing point $C_g U = e$. After a waiting period, a second pulse is applied which, in absence of decoherence, would make the qubit return to the ground state. Finally, after these manipulations, the read-out transistor is switched on while the box gate charge is moved away from $C_g U = e$. If decoherence occurs, the qubit will be left in the $n = 2$ charge state with a non-zero probability p_2 . Given its sensitivity and low back-action noise, the r.f.-SET should be able to detect this event with a signal-to-noise ratio better than unity. C_g and L_{tr} refer respectively to the capacitance and inductance of the impedance transformer. The times t_w and t_ϕ are the waiting time between the two pulses and the intrinsic decoherence time of the box, respectively.

randomness of tunnel events is dominated by phase-scrambling processes inside the island. But in the co-tunnelling regime, back-action noise is dominated by fluctuations of the island voltage associated with the measurement of the gate voltage itself and should therefore be more efficient. More work, experimental and theoretical, is needed to quantify this conjecture.

The superconducting SET can measure the charge on the gate with fully coherent carriers, the Cooper pairs. Zorin has analysed theoretically a version of this device shunted by a resistor and found that the noise energy could in this case exactly reach the quantum limit²⁹. However, the mode of operation seems to involve a more precise tuning of the device parameters than for the SET.

So far we have considered voltage amplifiers, that is, amplifiers with an input impedance that is large compared with the resistance quantum R_K . For several years already, SQUID devices have played the role of the SET in the realm of low-impedance amplifiers. There is in fact a duality relationship between the d.c.-SQUID and the SET¹. Whereas the SET is based on charge quantization of a metallic island surrounded by an insulator, the d.c.-SQUID is based on the quantization of flux in a superconducting loop. The sub-electron sensitivity of the SET has its analogue in the sub-flux quantum sensitivity of the SQUID. Although it has been known for many years that a SQUID could operate at the quantum limit, this has been achieved only recently in the r.f. domain. Andre *et al.* have shown that a

SQUID with a microstrip input line could achieve $k_B T_N \approx 5\hbar\omega$ at 438 MHz (ref. 30).

Finally, we should mention a special class of linear amplifiers: the mixers based on photon-assisted tunnelling in superconductor–insulator–superconductor (SIS) junctions. They convert a signal at several hundred GHz into a signal at several GHz. Even though the absolute power of the output signal is weaker than the power of the input signal, these devices have a large ‘photon number’ gain. Furthermore, their noise closely approaches the quantum limit: $E_N = 0.6\hbar\omega$ (ref. 31). Closely related to SIS mixers are Josephson parametric amplifiers, which have been operated at the quantum limit³². However, these devices do not have the advantage of the SET and the SQUID to amplify at the same time both a.f. and r.f. signals, a useful feature for tuning and trouble-shooting.

Measuring the state of a two-level system

In the past few years there has been much interest in the possibility of realizing a quantum computer⁶. Although the more advanced experiments in this field are taking place in quantum optics systems^{33,34}, several implementations of quantum bits and quantum gates in solid-state systems have been proposed^{35–38}. We focus here on a charge qubit which is now well understood experimentally^{12,13} and theoretically^{39,40}: the Cooper-pair box (Fig. 7). It consists of a superconducting island with Coulomb energy E_C connected to a superconducting charge reservoir by means of a Josephson tunnel junction with Josephson coupling energy E_J . The island is influenced electrostatically through a gate capacitor C_g connected to voltage source (Fig. 7a). If the conditions $k_B T \ll E_J \ll E_C < \Delta$ are realized, where Δ is the energy gap of the superconductor, then the island will have only an even number n of excess electrons. This only degree of freedom will be a good quantum number, except at the avoided crossings of the charge levels (Fig. 7b,c). A recent experiment has shown that such a solid-state qubit was able to display several Rabi oscillations when stimulated by a microwave pulse¹³. However, the measurement scheme in this experiment involved a probe tunnel junction which observed the qubit continuously, but weakly. Such a continuous observation dephases the qubit and prevents a measurement of its intrinsic decoherence time. It is therefore desirable to read-out the qubit only at the end of the coherent manipulation and evolution period (Fig. 8). An obvious candidate for the read-out is the SET. In a version with r.f. bias⁴, the transistor can be turned on and off very quickly, the turn-on time being set by the damping time of the tank circuit. This latter is of the order of several tens of nanoseconds, a time supposedly shorter than the intrinsic decoherence time of the box, estimated to be longer than 1 μ s. The question therefore arises as to whether the SET is sufficiently sensitive to detect the state of the box with a signal-to-noise ratio allowing the study of decoherence mechanisms.

Measurement time and dephasing rate

An important concept is the measurement time T_m of the read-out SET¹⁴, which is defined as the time needed for discriminating between the two charge states of the box differing by a Cooper pair, assumed to be good eigenstates, with a signal-to-noise ratio of 1. Island box charge is a good quantum number when $E_J/E_{cl} \ll 1$, where the electrostatic energy $E_{cl} = 2e^2(1 - C_g U/e)/C_{tot}$ of the box is controlled by the voltage U (here C_g and C_{tot} refer to the box gate capacitance and total island capacitance). In practice, the charge measurement is made just after U is varied away from the crossing point (Fig. 8b). The circuit of Fig. 8a shows that T_m is closely related to the voltage noise of the SET, with $T_m = 4S_V/(2e/C_{tot})^2$. We thus find that $T_m = (\delta Q/e)^2(C_{tot}/C_{in})^2$, where C_{in}/C_{tot} is the charge-coupling coefficient.

It is interesting to note that the back-action current noise dephases the charge states relative to one another only if charge is a

good quantum number. The general expression for the dephasing rate is

$$\Gamma_{\phi} = \left(\frac{e}{\hbar} \right)^2 \frac{E_{\text{el}}^2}{E_j^2 + E_{\text{el}}^2} S_V^{\text{box}}(\omega=0)$$

where $S_V^{\text{box}}(\omega) = S_I(\omega)/C_{\text{tot}}\omega^2$ is the fluctuation spectral density of the box island voltage induced by the SET current noise. In the regime $E_j/E_{\text{el}} \ll 1$, we find that the product $T_m \Gamma_{\phi} = 2[E_N(\omega)/\hbar\omega]^2_{\omega=0}$ involves only the noise energy and is of order unity. Once again, this is another close approach to the quantum limit as the minimum value for $T_m \Gamma_{\phi}$ is $\frac{1}{2}$ (Table 1).

So in principle it seems that the SET could acquire charge with any given large signal-to-noise ratio by measuring the box for a long enough time. However, any coupling between charge states will induce transitions and will corrupt the measurement. The transition rate between charge states due to the current noise is given by

$$\Gamma_1 = \left(\frac{e}{\hbar} \right)^2 \frac{E_j^2}{E_j^2 + E_{\text{el}}^2} S_V^{\text{box}}(\omega=\Omega)$$

where $\Omega = \sqrt{E_j^2 + E_{\text{el}}^2}/\hbar$ is the transition frequency between charge states.

We thus find that the signal-to-noise ratio including the effect of transitions between charge states is

$$S/N = (\Gamma_1 T_m)^{-1/2} = \frac{1}{(E_N/\hbar\omega)_{\omega=0}} \left(\frac{2E_j^2}{E_j^2 + E_{\text{el}}^2} \right)^{-1/2}$$

In practice, a signal-to-noise ratio significantly greater than 1 seems possible with an optimized r.f.-SET in a single charge measurement, even when taking into account the 50% reduction in sensitivity of the SET in the r.f.-bias mode⁴¹.

Towards single-photon sensitivity in the r.f. domain

To summarize, the SET transistor is a charge amplifying device whose sensitivity in the r.f. domain is limited in principle only by a well understood process, quantum shot noise. This property displays a marked contrast with a FET. It should be possible in the near future to show experimentally that the noise energy of the SET approaches the quantum limit within a factor of order unity. Although the SET will not be 100% efficient in acquiring charge information with only the minimal back-action noise required by quantum mechanics, it should still be able to read-out a charge qubit in a single-shot measurement. In more sophisticated devices using Cooper-pair tunnelling or co-tunnelling processes, the quantum limit could be approached even more closely. Furthermore, by transposing in the r.f. domain the manipulations of the quantum signal that are now performed routinely in experiments in cavity quantum electrodynamics⁴², one could use these ultimate amplifiers for measuring signals consisting of a single photon, without even destroying the photon. □

1. Averin, D. V. & Likharev, K. K. Coulomb blockade of tunneling and coherent oscillations in small tunnel junctions. *J. Low Temp. Phys.* **62**, 345–372 (1986).
2. Fulton, T. A. & Dolan, G. J. Observation of single-electron charging effects in small tunnel junctions. *Phys. Rev. Lett.* **59**, 109–112 (1987).
3. Meirav U., Kastner, M. A. & Wind, S. J. Single-electron charging and periodic conductance resonances in GaAs nanostructures. *Phys. Rev. Lett.* **65**, 771–774 (1990).
4. Schoelkopf, R. J., Wahlgren, P., Kozhevnikov, A. A., Delsing, P. & Prober, D. The radio-frequency single-

- electron transistor (RF-SET): a fast and ultrasensitive electrometer. *Science* **280**, 1238–1242 (1998).
5. Mather, J. C. Super photon counters. *Nature* **401**, 654–655 (1999).
6. Steane, A. Quantum computation. *Rep. Prog. Phys.* **61**, 117 (1998).
7. Korotkov, A. N., Averin, D., Likharev, K. K. & Vasenko, S. A. in *Single-Electron Tunneling and Mesoscopic Physics* (eds Koch, H. & Lubbick, H.) 45–59 (Springer, Berlin, 1992).
8. Mar, D. J., Westervelt R. & Hopkins, P. F. Cryogenic field-effect transistor with single electronic charge sensitivity. *Appl. Phys. Lett.* **64**, 631–633 (1994).
9. Keller, M. W., Eichenberger, A., Martinis, J. M. & Zimmerman, N. M. A capacitance standard based on counting electrons. *Science* **285**, 1706–1709 (1999).
10. Schoelkopf, R. J., Moseley, S. H., Stahle, C. M., Wahlgreen, P. & Delsing, P. A concept for a submillimeter-wave single-photon counter? *IEEE Trans. Appl. Supercond.* **9**, 2935–2938 (1999).
11. Komiya, S., Astafiev, O., Antonov, V., Kutsuwa, T. & Hirai, H. A single-photon detector in the far-infrared range. *Nature* **403**, 405–407 (2000).
12. Bouchiat, V., Vion, D., Joyez, P., Esteve, D. & Devoret, M. H. Quantum coherence with a single Cooper pair. *Phys. Scr.* **T76**, 165–170 (1998).
13. Nakamura, Y., Pashkin, Yu. A. & Tsai, J. S. Coherent control of macroscopic quantum states in a single-Cooper-pair box. *Nature* **398**, 786–788 (1999).
14. Shnirman, A. & Schoen, G. Quantum measurements performed with a single-electron transistor. *Phys. Rev. B* **57**, 15400–15407 (1997).
15. Naveh, Y. & Likharev, K. K. Modeling of 10-nm-scale ballistic MOSFET's. *IEEE Electron Devices Lett.* **21**, 242–244 (2000).
16. Devoret, M. H. & Grabert, H. in *Single Charge Tunneling* (eds Grabert, H. & Devoret, M. H.) 1–19 (Plenum, New York, 1992).
17. Grabert, H. Charge fluctuations in the single-electron box: perturbation expansion in the tunneling conductance. *Phys. Rev.* **50**, 17364–17377 (1994).
18. Schoeller, H. & Schoen, G. Mesoscopic quantum transport: resonant tunneling in the presence of a strong Coulomb interaction. *Phys. Rev. B* **50**, 18436–18442 (1994).
19. Shirakashi, J., Matsumoto, K., Miura, N. & Konagai, N. Single-electron charging effects in Nb/Nb oxide-based single-electron transistors at room temperature. *Appl. Phys. Lett.* **72**, 1893–1895 (1998).
20. Zhuang, L., Guo, L. & Chou, S. Y. Silicon single-electron quantum-dot transistor switch operating at room temperature. *Appl. Phys. Lett.* **72**, 1205–1207 (1998).
21. Pashkin, Yu. A., Nakamura, Y. & Tsai, J. S. Room-temperature Al single-electron transistor made by electron-beam lithography. *Appl. Phys. Lett.* **76**, 2256–2258 (2000).
22. Wolf, H. et al. Investigation of the offset charge noise in single electron tunneling devices. *IEEE Trans. Instrum. Measurement* **46**, 303–306 (1997).
23. Caves, C. M., Thorne, K. S., Drever, W. P., Sandberg, V. D. & Zimmermann, N. On the measurement of a weak classical force coupled to a quantum-mechanical oscillator. I. Issues of principle. *Rev. Mod. Phys.* **52**, 341–392 (1980).
24. Caves, C. M. Quantum limits on noise in linear amplifiers. *Phys. Rev. D* **26**, 1817–1839 (1982).
25. Braginsky, V. B. & Khalili, F. Ya. *Quantum Measurement* (Cambridge Univ. Press, 1992).
26. Pospieszalski, M. W. & Wollack, E. J. in *Proceedings of 2nd ESA Workshop on Millimetre Wave Technology and Applications (WPP-149)* 221–226 (ESA, Paris, 1998).
27. Gurvitz, S. A. Measurements with a noninvasive detector and dephasing mechanism. *Phys. Rev. B* **56**, 15215–15223 (1997).
28. Korotkov, A. N. Preprint cond-mat/0003225 at <http://xxx.lanl.gov> (2000).
29. Zorin, A. B. Quantum-limited electrometer based on single Cooper pair tunneling. *Phys. Rev. Lett.* **76**, 4408–4411 (1996).
30. André, M.-O., Mück, M., Clarke, J., Gail, J. & Heiden, C. Radio-frequency amplifier with tenth-kelvin noise temperature based on microstrip direct current superconducting quantum interference device. *Appl. Phys. Lett.* **75**, 698–700 (1999).
31. Mears, C. A. et al. Quantum-limited heterodyne detection of millimeter waves using superconducting tantalum tunnel junctions. *Appl. Phys. Lett.* **57**, 2487–2489 (1990).
32. Movshovich, R. et al. Observation of zero-point noise squeezing via a Josephson parametric amplifier. *Phys. Rev. Lett.* **65**, 1419–1422 (1990).
33. Turchette, Q. A., Hood, C. J., Lange, W., Mabuchi, H. & Kimble, H. J. Measurement of conditional phase shifts for quantum logic. *Phys. Rev. Lett.* **75**, 4710–4713 (1995).
34. Monroe, C. et al. Resolved-sideband Raman cooling of a bound atom to the 3D zero-point energy. *Phys. Rev. Lett.* **75**, 4011–4014 (1995).
35. Bocko, M. E., Herr, A. M. & Feldman, M. F. Prospects for quantum coherent computation using superconducting electronics. *IEEE Trans. Appl. Supercond.* **7**, 3638–3641 (1997).
36. Kane, B. E. A silicon-based nuclear spin quantum computer. *Nature* **393**, 133–137 (1998).
37. Loss, D. & DiVincenzo, D. P. Quantum computation with quantum dots. *Phys. Rev. A* **57**, 120–126 (1998).
38. Mooij, J. E. et al. Josephson persistent-current qubit. *Science* **285**, 1036–1039 (1999).
39. Averin, D. V. Adiabatic quantum computation with Cooper pairs. *Solid State Commun.* **105**, 657–659 (1998).
40. Makhlin, Yu., Schoen, G. & Shnirman, A. Josephson-junction qubits with controlled couplings. *Nature* **398**, 786–789 (1999).
41. Korotkov, A. N. & Paalanen, M. A. Charge sensitivity of radio frequency single-electron transistor. *Appl. Phys. Lett.* **74**, 4052–4054 (1999).
42. Noguees, C. et al. Seeing a single photon without destroying it. *Nature* **400**, 239–242 (1999).

Acknowledgements

This work was supported in part by the National Security Agency (NSA), the Advanced Research and Development Activity (ARDA) and the Army Research Office (ARO). One of us (M.H.D.) acknowledges support from Commissariat à l'Energie Atomique (CEA). We thank D. Averin, J. Clarke, P. Delsing, D. Esteve, A. Korotkov, K. Likharev, H. Mooij, D. Prober, G. Schoen and E. Wollack for helpful discussions and communications of results prior to publication.

Ultimate physical limits to computation

Seth Lloyd

d'Arbelloff Laboratory for Information Systems and Technology, MIT Department of Mechanical Engineering, Massachusetts Institute of Technology 3-160, Cambridge, Massachusetts 02139, USA (slloyd@mit.edu)

Computers are physical systems: the laws of physics dictate what they can and cannot do. In particular, the speed with which a physical device can process information is limited by its energy and the amount of information that it can process is limited by the number of degrees of freedom it possesses. Here I explore the physical limits of computation as determined by the speed of light c , the quantum scale $\hbar$ and the gravitational constant G . As an example, I put quantitative bounds to the computational power of an 'ultimate laptop' with a mass of one kilogram confined to a volume of one litre.

Over the past half century, the amount of information that computers are capable of processing and the rate at which they process it has doubled every 18 months, a phenomenon known as Moore's law. A variety of technologies — most recently, integrated circuits — have enabled this exponential increase in information processing power. But there is no particular reason why Moore's law should continue to hold: it is a law of human ingenuity, not of nature. At some point, Moore's law will break down. The question is, when?

The answer to this question will be found by applying the laws of physics to the process of computation^{1–85}. Extrapolation of current exponential improvements over two more decades would result in computers that process information at the scale of individual atoms. Although an Avogadro-scale computer that can act on 10^{23} bits might seem implausible, prototype quantum computers that store and process information on individual atoms have already been demonstrated^{64,65,76–80}. Existing quantum computers may be small and simple, and able to perform only a few hundred operations on fewer than ten quantum bits or 'qubits', but the fact that they work at all indicates that there is nothing in the laws of physics that forbids the construction of an Avogadro-scale computer.

The purpose of this article is to determine just what limits the laws of physics place on the power of computers. At first, this might seem a futile task: because we do not know the technologies by which computers 1,000, 100, or even 10 years in the future will be constructed, how can we determine the physical limits of those technologies? In fact, I will show that a great deal can be determined concerning the ultimate physical limits of computation simply from knowledge of the speed of light, $c = 2.9979 \times 10^8 \text{ m s}^{-1}$, Planck's reduced constant, $\hbar = h/2\pi = 1.0545 \times 10^{-34} \text{ J s}$, and the gravitational constant, $G = 6.673 \times 10^{-11} \text{ m}^3 \text{ kg}^{-1} \text{ s}^{-2}$. Boltzmann's constant, $k_B = 1.3805 \times 10^{-23} \text{ J K}^{-1}$, will also be crucial in translating between computational quantities such as memory space and operations per bit per second, and thermodynamic quantities such as entropy and temperature. In addition to reviewing previous work on how physics limits the speed and memory of computers, I present results — which are new except as noted — of the derivation of the ultimate speed limit to computation, of trade-offs between memory and speed, and of the analysis of the behaviour of computers at physical extremes of high temperatures and densities.

Before presenting methods for calculating these limits, it is important to note that there is no guarantee that these limits will ever be attained, no matter how ingenious

computer designers become. Some extreme cases such as the black-hole computer described below are likely to prove extremely difficult or impossible to realize. Human ingenuity has proved great in the past, however, and before writing off physical limits as unattainable, we should realize that certain of these limits have already been attained within a circumscribed context in the construction of working quantum computers. The discussion below will note obstacles that must be sidestepped or overcome before various limits can be attained.

Energy limits speed of computation

To explore the physical limits of computation, let us calculate the ultimate computational capacity of a computer with a mass of 1 kg occupying a volume of 1 litre, which is roughly the size of a conventional laptop computer. Such a computer, operating at the limits of speed and memory space allowed by physics, will be called the 'ultimate laptop' (Fig. 1).

First, ask what limits the laws of physics place on the speed of such a device. As I will now show, to perform an elementary logical operation in time Δt requires an average amount of energy $E \geq \pi \hbar / 2 \Delta t$. As a consequence, a system with average energy E can perform a maximum of $2E / \pi \hbar$ logical operations per second. A 1-kg computer has average energy $E = mc^2 = 8.9874 \times 10^{16} \text{ J}$. Accordingly, the ultimate laptop can perform a maximum of 5.4258×10^{50} operations per second.

Maximum speed per logical operation

For the sake of convenience, the ultimate laptop will be taken to be a digital computer. Computers that operate on non-binary or continuous variables obey similar limits to those that will be derived here. A digital computer performs computation by representing information in the terms of binary digits or bits, which can take the value 0 or 1, and then processes that information by performing simple logical operations such as AND, NOT and FANOUT. The operation, AND, for instance, takes two binary inputs X and Y and returns the output 1 if and only if both X and Y are 1; otherwise it returns the output 0. Similarly, NOT takes a single binary input X and returns the output 1 if $X = 0$ and 0 if $X = 1$. FANOUT takes a single binary input X and returns two binary outputs, each equal to X . Any boolean function can be constructed by repeated application of AND, NOT and FANOUT. A set of operations that allows the construction of arbitrary boolean functions is called universal. The actual physical device that performs a logical operation is called a logic gate.

How fast can a digital computer perform a logical operation? During such an operation, the bits in the computer on



Figure 1 The ultimate laptop. The 'ultimate laptop' is a computer with a mass of 1 kg and a volume of 1 l, operating at the fundamental limits of speed and memory capacity fixed by physics. The ultimate laptop performs $2mc^2/\pi\hbar = 5.4258 \times 10^{50}$ logical operations per second on $\sim 10^{31}$ bits. Although its computational machinery is in fact in a highly specified physical state with zero entropy, while it performs a computation that uses all its resources of energy and memory space it appears to an outside observer to be in a thermal state at $\sim 10^9$ degrees Kelvin. The ultimate laptop looks like a small piece of the Big Bang.

which the operation is performed go from one state to another. The problem of how much energy is required for information processing was first investigated in the context of communications theory by Levitin^{11–16}, Bremermann^{17–19}, Beckenstein^{20–22} and others, who showed that the laws of quantum mechanics determine the maximum rate at which a system with spread in energy ΔE can move from one distinguishable state to another. In particular, the correct interpretation of the time–energy Heisenberg uncertainty principle $\Delta E \Delta t \geq \hbar$ is not that it takes time Δt to measure energy to an accuracy ΔE (a fallacy that was put to rest by Aharonov and Bohm^{23,24}), but rather that a quantum state with spread in energy ΔE takes time at least $\Delta t = \pi\hbar/2\Delta E$ to evolve to an orthogonal (and hence distinguishable) state^{23–26}. More recently, Margolus and Levitin^{15,16} extended this result to show that a quantum system with average energy E takes time at least $\Delta t = \pi\hbar/2E$ to evolve to an orthogonal state.

Performing quantum logic operations

As an example, consider the operation NOT performed on a qubit with logical states $|0\rangle$ and $|1\rangle$. (For readers unfamiliar with quantum mechanics, the 'bracket' notation $|\rangle$ signifies that whatever is contained in the bracket is a quantum-mechanical variable; $|0\rangle$ and $|1\rangle$ are vectors in a two-dimensional vector space over the complex numbers.) To flip the qubit, one can apply a potential $H = E_0|E_0\rangle\langle E_0| + E_1|E_1\rangle\langle E_1|$ with energy eigenstates $|E_0\rangle = (1/\sqrt{2})(|0\rangle + |1\rangle)$ and $|E_1\rangle = (1/\sqrt{2})(|0\rangle - |1\rangle)$. Because $|0\rangle = (1/\sqrt{2})(|E_0\rangle + |E_1\rangle)$ and $|1\rangle = (1/\sqrt{2})(|E_0\rangle - |E_1\rangle)$, each logical state $|0\rangle, |1\rangle$ has spread in energy $\Delta E = (E_1 - E_0)/2$. It is easy to verify that after a length of time $\Delta t = \pi\hbar/2\Delta E$ the qubit evolves so that $|0\rangle \rightarrow |1\rangle$ and $|1\rangle \rightarrow |0\rangle$. That is, applying the potential effects a NOT operation in a time that attains the limit given by quantum mechanics. Note that the average energy E of the qubit in the course of the logical operation is $\langle 0|H|0\rangle = \langle 1|H|1\rangle = (E_0 + E_1)/2 = E_0 + \Delta E$. Taking the ground-state energy $E_0 = 0$ gives $E = \Delta E$. So the amount of time it takes to perform a NOT operation can also be written as $\Delta t = \pi\hbar/2E$. It is straightforward to show^{15,16} that no quantum system with average energy E can move to an orthogonal state in a time less than Δt . That is, the speed with which a logical operation can be performed is limited not only by the spread in energy, but also by the average energy. This result will prove to be a key component in deriving the speed limit for the ultimate laptop.

AND and FANOUT can be enacted in a way that is analogous to the NOT operation. A simple way to perform these operations in a quantum-mechanical context is to enact a so-called Toffoli or controlled-controlled-NOT operation³¹. This operation takes three binary inputs, X , Y and Z , and returns three outputs, X' , Y' and Z' .

The first two inputs pass through unchanged, that is, $X' = X$, $Y' = Y$. The third input passes through unchanged unless both X and Y are 1, in which case it is flipped. This is universal in the sense that suitable choices of inputs allow the construction of AND, NOT and FANOUT. When the third input is set to zero, $Z = 0$, then the third output is the AND of the first two: $Z' = X \text{ AND } Y$. So AND can be constructed. When the first two inputs are 1, $X = Y = 1$, the third output is the NOT of the third input, $Z' = \text{NOT } Z$. Finally, when the second input is set to 1, $Y = 1$, and the third to zero, $Z = 0$, the first and third output are the FANOUT of the first input, $X' = X$, $Z' = X$. So arbitrary boolean functions can be constructed from the Toffoli operation alone.

By embedding a controlled-controlled-NOT gate in a quantum context, it is straightforward to see that AND and FANOUT, like NOT, can be performed at a rate $2E/\pi\hbar$ times per second, where E is the average energy of the logic gate that performs the operation. More complicated logic operations that cycle through a larger number of quantum states (such as those on non-binary or continuous quantum variables) can be performed at a rate $E/\pi\hbar$ — half as fast as the simpler operations^{15,16}. Existing quantum logic gates in optical-atomic and nuclear magnetic resonance (NMR) quantum computers actually attain this limit. In the case of NOT, E is the average energy of interaction of the qubit's dipole moment (electric dipole for optic-atomic qubits and nuclear magnetic dipole for NMR qubits) with the applied electromagnetic field. In the case of multi-qubit operations such as the Toffoli operation, or the simpler two-bit controlled-NOT operation, which flips the second bit if and only if the first bit is 1, E is the average energy in the interaction between the physical systems that register the qubits.

Ultimate limits to speed of computation

We are now in a position to derive the first physical limit to computation, that of energy. Suppose that one has a certain amount of energy E to allocate to the logic gates of a computer. The more energy one allocates to a gate, the faster it can perform a logic operation. The total number of logic operations performed per second is equal to the sum over all logic gates of the operations per second per gate. That is, a computer can perform no more than

$$\sum_{\ell} 1/\Delta t_{\ell} \leq \sum_{\ell} 2E_{\ell}/\pi\hbar = 2E/\pi\hbar$$

operations per second. In other words, the rate at which a computer can compute is limited by its energy. (Similar limits have been proposed by Bremermann in the context of the minimum energy

required to communicate a bit^{17–19}, although these limits have been criticized for misinterpreting the energy–time uncertainty relation²¹, and for failing to take into account the role of degeneracy of energy eigenvalues^{13,14} and the role of nonlinearity in communications^{7–9}.) Applying this result to a 1-kg computer with energy $E = mc^2 = 8.9874 \times 10^{16}$ J shows that the ultimate laptop can perform a maximum of 5.4258×10^{50} operations per second.

Parallel and serial operation

An interesting feature of this limit is that it is independent of computer architecture. It might have been thought that a computer could be speeded up by parallelization, that is, by taking the energy and dividing it up among many subsystems computing in parallel. But this is not the case. If the energy E is spread among N logic gates, each one operates at a rate $2E/\pi\hbar N$, and the total number of operations per second, $N2E/\pi\hbar N = 2E/\pi\hbar$, remains the same. If the energy is allocated to fewer logic gates (a more serial operation), the rate $1/\Delta t_\ell$ at which they operate and the spread in energy per gate ΔE_ℓ increase. If the energy is allocated to more logic gates (a more parallel operation) then the rate at which they operate and the spread in energy per gate decrease. Note that in this parallel case, the overall spread in energy of the computer as a whole is considerably smaller than the average energy: in general $\Delta E = \sqrt{\sum \Delta E_\ell^2} \approx \sqrt{N\Delta E_\ell^2}$ whereas $E = \sum E_\ell \approx NE_\ell$. Parallelization can help perform certain computations more efficiently, but it does not alter the total number of operations per second. As I will show below, the degree of parallelizability of the computation to be performed determines the most efficient distribution of energy among the parts of the computer. Computers in which energy is relatively evenly distributed over a larger volume are better suited for performing parallel computations. More compact computers and computers with an uneven distribution of energy are better for performing serial computations.

Comparison with existing computers

Conventional laptops operate much more slowly than the ultimate laptop. There are two reasons for this inefficiency. First, most of the energy is locked up in the mass of the particles of which the computer is constructed, leaving only an infinitesimal fraction for performing logic. Second, a conventional computer uses many degrees of freedom (billions and billions of electrons) for registering a single bit. From the physical perspective, such a computer operates in a highly redundant fashion. There are, however, good technological reasons for such redundancy, with conventional designs depending on it for reliability and manufacturability. But in the present discussion, the subject is not what computers are but what they might be, and in this context the laws of physics do not require redundancy to perform logical operations — recently constructed quantum microcomputers use one quantum degree of freedom for each bit and operate at the Heisenberg limit $\Delta t = \pi\hbar/2\Delta E$ for the time needed to flip a bit^{64,65,76–80}. Redundancy is, however, required for error correction, as will be discussed below.

In sum, quantum mechanics provides a simple answer to the question of how fast information can be processed using a given amount of energy. Now it will be shown that thermodynamics and statistical mechanics provide a fundamental limit to how many bits of information can be processed using a given amount of energy confined to a given volume. Available energy necessarily limits the rate at which a computer can process information. Similarly, the maximum entropy of a physical system determines the amount of information it can process. Energy limits speed. Entropy limits memory.

Entropy limits memory space

The amount of information that a physical system can store and process is related to the number of distinct physical states that are accessible to the system. A collection of m two-state systems has 2^m accessible states and can register m bits of information. In general, a system with N accessible states can register $\log_2 N$ bits of information. But it has been known for more than a century that the number of accessible states of a physical system, W , is related to its

thermodynamic entropy by the formula $S = k_B \ln W$, where k_B is Boltzmann's constant. (Although this formula is inscribed on Boltzmann's tomb, it is attributed originally to Planck; before the turn of the century, k_B was often known as Planck's constant.)

The amount of information that can be registered by a physical system is $I = S(E)/k_B \ln 2$, where $S(E)$ is the thermodynamic entropy of a system with expectation value for the energy E . Combining this formula with the formula $2E/\pi\hbar$ for the number of logical operations that can be performed per second, we see that when it is using all its memory, the number of operations per bit per second that our ultimate laptop can perform is $k_B 2 \ln(2) E / \pi\hbar S \propto k_B T / \hbar$, where $T = (\partial S / \partial E)^{-1}$ is the temperature of 1 kg of matter in a maximum entropy in a volume of 1 l. The entropy governs the amount of information the system can register and the temperature governs the number of operations per bit per second that it can perform.

Because thermodynamic entropy effectively counts the number of bits available to a physical system, the following derivation of the memory space available to the ultimate laptop is based on a thermodynamic treatment of 1 kg of matter confined to a volume of 1 l in a maximum entropy state. Throughout this derivation, it is important to remember that although the memory space available to the computer is given by the entropy of its thermal equilibrium state, the actual state of the ultimate laptop as it performs a computation is completely determined, so that its entropy remains always equal to zero. As above, I assume that we have complete control over the actual state of the ultimate laptop, and are able to guide it through its logical steps while insulating it from all uncontrolled degrees of freedom. But as the following discussion will make clear, such complete control will be difficult to attain (see Box 1).

Entropy, energy and temperature

To calculate the number of operations per second that could be performed by our ultimate laptop, I assume that the expectation value of the energy is E . Accordingly, the total number of bits of memory space available to the computer is $S(E, V)/k_B \ln 2$ where $S(E, V)$ is the thermodynamic entropy of a system with expectation value of the energy E confined to volume V . The entropy of a closed system is usually given by the so-called microcanonical ensemble, which fixes both the average energy and the spread in energy ΔE , and assigns equal probability to all states of the system within a range $[E, E + \Delta E]$. In the case of the ultimate laptop, however, I wish to fix only the average energy, while letting the spread in energy vary according to whether the computer is to be more serial (fewer, faster gates, with larger spread in energy) or parallel (more, slower gates, with smaller spread in energy). Accordingly, the ensemble that should be used to calculate the thermodynamic entropy and the memory space available is the canonical ensemble, which maximizes S for fixed average energy with no constraint on the spread in energy ΔE . The canonical ensemble shows how many bits of memory are available for all possible ways of programming the computer while keeping its average energy equal to E . In any given computation with average energy E , the ultimate laptop will be in a pure state with some fixed spread of energy, and will explore only a small fraction of its memory space.

In the canonical ensemble, a state with energy E_i has probability $p_i = (1/Z(T))e^{-E_i/k_B T}$ where $Z(T) = \sum_i e^{-E_i/k_B T}$ is the partition function, and the temperature T is chosen so that $E = \sum_i p_i E_i$. The entropy is $S = -k_B \sum_i p_i \ln p_i = E/T + k_B \ln Z$. The number of bits of memory space available to the computer is $S/k_B \ln 2$. The difference between the entropy as calculated using the canonical ensemble and that calculated using the microcanonical ensemble is minimal. But there is some subtlety involved in using the canonical ensemble rather than the more traditional microcanonical ensemble. The canonical ensemble is normally used for open systems that interact with a thermal bath at temperature T . In the case of the ultimate laptop, however, it is applied to a closed system to find the maximum entropy given a fixed expectation value for the energy. As a result, the temperature $T = (\partial S / \partial E)^{-1}$ has a somewhat different role in the context of physical limits of computation than it does in the case of an ordinary

Box 1

The role of thermodynamics in computation

The fact that entropy and information are intimately linked has been known since Maxwell introduced his famous ‘demon’ well over a century ago¹. Maxwell’s demon is a hypothetical being that uses its information-processing ability to reduce the entropy of a gas. The first results in the physics of information processing were derived in attempts to understand how Maxwell’s demon could function^{1–4}. The role of thermodynamics in computation has been examined repeatedly over the past half century. In the 1950s, von Neumann¹⁰ speculated that each logical operation performed in a computer at temperature T must dissipate energy $k_B T \ln 2$, thereby increasing entropy by $k_B \ln 2$. This speculation proved to be false. The precise, correct statement of the role of entropy in computation was attributed to Landauer⁵, who showed that reversible, that is, one-to-one, logical operations such as NOT can be performed, in principle, without dissipation, but that irreversible, many-to-one operations such as AND or ERASE require dissipation of at least $k_B \ln 2$ for each bit of information lost. (ERASE is a one-bit logical operation that takes a bit, 0 or 1, and restores it to 0.) The argument behind Landauer’s principle can be readily understood³⁷. Essentially, the one-to-one dynamics of hamiltonian systems implies that when a bit is erased the information that it contains has to go somewhere. If the information goes into observable degrees of freedom of the computer, such as another bit, then it has not been erased but merely moved; but if it goes into unobservable degrees of freedom such as the microscopic motion of molecules it results in an increase of entropy of at least $k_B \ln 2$.

In 1973, Bennett^{28–30} showed that all computations could be performed using only reversible logical operations. Consequently, by Landauer’s principle, computation does not require dissipation. (Earlier work by Lecerf²⁷ had anticipated the possibility of reversible computation, but not its physical implications. Reversible computation was discovered independently by Fredkin and Toffoli³¹.) The energy used to perform a logical operation can be ‘borrowed’ from a store of free energy such as a battery, ‘invested’ in the logic gate that performs the operation, and returned to storage after the operation has been performed, with a net ‘profit’ in the form of processed information. Electronic circuits based on reversible logic have been built and exhibit considerable reductions in dissipation over conventional reversible circuits^{33–35}.

Under many circumstances it may prove useful to perform irreversible operations such as erasure. If our ultimate laptop is subject to an error rate of ϵ bits per second, for example, then error-correcting codes can be used to detect those errors and reject them to the environment at a dissipative cost of $\epsilon k_B T_E \ln 2$ J s^{−1}, where T_E is the temperature of the environment. ($k_B T \ln 2$ is the minimal amount of energy required to send a bit down an information channel with noise temperature T (ref. 14).) Such error-correcting routines in our ultimate computer function as working analogues of Maxwell’s demon, getting information and using it to reduce entropy at an exchange rate of $k_B T \ln 2$ joules per bit. In principle, computation does not require dissipation. In practice, however, any computer — even our ultimate laptop — will dissipate energy.

The ultimate laptop must reject errors to the environment at a high rate to maintain reliable operation. To estimate the rate at which it can reject errors to the environment, assume that the computer encodes erroneous bits in the form of black-body radiation at the characteristic temperature 5.87×10^8 K of the computer’s memory²¹. The Stefan–Boltzmann law for black-body radiation then implies that the number of bits per unit area than can be sent out to the environment is $B = \pi^2 k_B^3 T^3 / 60 \ln(2) \hbar^3 c^2 = 7.195 \times 10^{42}$ bits per square meter per second. As the ultimate laptop has a surface area of 10^{-2} m² and is performing $\sim 10^{50}$ operations per second, it must have an error rate of less than 10^{-10} per operation in order to avoid over-heating. Even if it achieves such an error rate, it must have an energy throughput (free energy in and thermal energy out) of 4.04×10^{26} W — turning over its own resting mass energy of $mc^2 \approx 10^{17}$ J in a nanosecond! The thermal load of correcting large numbers of errors clearly indicates the necessity of operating at a slower speed than the maximum allowed by the laws of physics.

thermodynamic system interacting with a thermal bath. Integrating the relationship $T = (\partial S / \partial E)^{-1}$ over E yields $T = CE/S$, where C is a constant of order unity (for example, $C = 4/3$ for black-body radiation, $C = 3/2$ for an ideal gas, and $C = 1/2$ for a black hole). Accordingly, the temperature governs the number of operations per bit per second, $k_B \ln(2) E / \hbar S \approx k_B T / \hbar$, that a system can perform. As I will show later, the relationship between temperature and operations per bit per second is useful in investigating computation under extreme physical conditions.

Calculating the maximum memory space

To calculate exactly the maximum entropy for a kilogram of matter in a litre volume would require complete knowledge of the dynamics of elementary particles, quantum gravity, and so on. Although we do not possess such knowledge, the maximum entropy can readily be estimated by a method reminiscent of that used to calculate thermodynamic quantities in the early Universe⁸⁶. The idea is simple: model the volume occupied by the computer as a collection of modes of elementary particles with total average energy E . The maximum entropy is obtained by calculating the canonical ensemble over the modes. Here, I supply a simple derivation of the maximum memory space available to the ultimate laptop. A more detailed discussion of how to calculate the maximum amount of information that can be stored in a physical system can be found in the work of Bekenstein^{19–21}.

For this calculation, assume that the only conserved quantities other than the computer’s energy are angular momentum and electric charge, which I take to be zero. (One might also ask that the number of baryons be conserved, but as will be seen below, one of the processes that could take place within the computer is black-hole formation and evaporation, which does not conserve baryon number.) At a particular temperature T , the entropy is dominated by

the contributions from particles with mass less than $k_B T / 2c^2$. The ℓ th such species of particle contributes energy $E = r_\ell \pi^2 V (k_B T)^4 / 30 \hbar^3 c^3$ and entropy $S = 2 r_\ell k_B \pi^2 V (k_B T)^3 / 45 \hbar^3 c^3 = 4E/3T$, where r_ℓ is equal to the number of particles/antiparticles in the species (that is, 1 for photons, 2 for electrons/positrons) multiplied by the number of polarizations (2 for photons, 2 for electrons/positrons) multiplied by a factor that reflects particle statistics (1 for bosons, 7/8 for fermions). As the formula for S in terms of T shows, each species contributes $(2\pi)^5 r_\ell / 90 \ln 2 \approx 10^2$ bits of memory space per cubic thermal wavelength λ_T^3 , where $\lambda_T = 2\pi \hbar c / k_B T$. Re-expressing the formula for entropy as a function of energy, the estimate for the maximum entropy is

$$S = (4/3) k_B (\pi^2 r V / 30 \hbar^3 c^3)^{1/4} E^{3/4} = k_B \ln(2) I$$

where $r = \sum_\ell r_\ell$. Note that S depends only insensitively on the total number of species with mass less than $k_B T / 2c^2$.

A lower bound on the entropy can be obtained by assuming that energy and entropy are dominated by black-body radiation consisting of photons. In this case, $r = 2$, and for a 1-kg computer confined to a volume of 1 l we have $k_B T = 8.10 \times 10^{-15}$ J, or $T = 5.87 \times 10^8$ K. The entropy is $S = 2.04 \times 10^8$ J K^{−1}, which corresponds to an amount of available memory space $I = S / k_B \ln 2 = 2.13 \times 10^{31}$ bits. When the ultimate laptop is using all its memory space it can perform $2 \ln(2) k_B E / \pi \hbar S = 3 \ln(2) k_B T / 2 \pi \hbar \approx 10^{19}$ operations per bit per second. As the number of operations per second $2E / \pi \hbar$ is independent of the number of bits available, the number of operations per bit per second can be increased by using a smaller number of bits. In keeping with the prescription that the ultimate laptop operates at the absolute limits given by physics, in what follows I assume that all available bits are used.

This estimate for the maximum entropy could be improved (and slightly increased) by adding more species of massless particles (neutrinos and gravitons) and by taking into effect the presence of electrons and positrons. Note that $k_B T/2c^2 = 4.51 \times 10^{-32}$ kg, compared with the electron mass of 9.1×10^{-31} kg. That is, the ultimate laptop is close to a phase transition at which electrons and positrons are produced thermally. A more exact estimate of the maximum entropy and hence the available memory space would be straightforward to perform, but the details of such a calculation would detract from my general exposition, and could serve to alter S only slightly. S depends insensitively on the number of species of effectively massless particles: a change of r by a factor of 10,000 serves to increase S by only a factor of 10.

Comparison with current computers

The amount of information that can be stored by the ultimate laptop, $\sim 10^{31}$ bits, is much higher than the $\sim 10^{10}$ bits stored on current laptops. This is because conventional laptops use many degrees of freedom to store a bit whereas the ultimate laptop uses just one. There are considerable advantages to using many degrees of freedom to store information, stability and controllability being perhaps the most important. Indeed, as the above calculation indicates, to take full advantage of the memory space available, the ultimate laptop must turn all its matter into energy. A typical state of the ultimate laptop's memory looks like a plasma at a billion degrees Kelvin — like a thermonuclear explosion or a little piece of the Big Bang! Clearly, packaging issues alone make it unlikely that this limit can be obtained, even setting aside the difficulties of stability and control.

Even though the ultimate physical limit to how much information can be stored in a kilogram of matter in a litre volume is unlikely to be attained, it may nonetheless be possible to progress some way towards such bit densities. In other words, the ultimate limits to memory space may prove easier to approach than the ultimate limits to speed. Following Moore's law, the density of bits in a computer has gone down from approximately one per square centimetre 50 years ago to one per square micrometre today, an improvement of a factor of 10^8 . It is not inconceivable that a similar improvement is possible over the course of the next 50 years. In particular, there is no physical reason why it should not be possible to store one bit of information per atom. Indeed, existing NMR and ion-trap quantum computers already store information on individual nuclei and atoms (typically in the states of individual nuclear spins or in hyperfine atomic states). Solid-state NMR with high gradient fields or quantum optical techniques such as spectral hole-burning provide potential technologies for storing large quantities of information at the atomic scale. A kilogram of ordinary matter holds on the order of 10^{25} nuclei. If a substantial fraction of these nuclei can be made to register a bit, then we could get close to the ultimate physical limit of memory without having to resort to thermonuclear explosions. If, in addition, we make use of the natural electromagnetic interactions between nuclei and electrons in the matter to perform logical operations, we are limited to a rate of $\sim 10^{15}$ operations per bit per second, yielding an overall information processing rate of $\sim 10^{40}$ operations per second in ordinary matter. Although less than the $\sim 10^{51}$ operations per second in the ultimate laptop, the maximum information processing rate in 'ordinary matter' is still quite respectable. Of course, even though such an 'ordinary matter' ultimate computer need not operate at nuclear energy levels, other problems remain — for example, the high number of bits still indicates substantial input/output problems. At an input/output rate of 10^{12} bits per second, an Avogadro-scale computer with 10^{23} bits would take about 10,000 years to perform a serial read/write operation on the entire memory. Higher throughput and parallel input/output schemes are clearly required to take advantage of the entire memory space that physics makes available.

Size limits parallelization

Up until this point, I have assumed that the ultimate laptop occupies a volume of 1 l. The previous discussion, however, indicates that

Box 2

Can a black hole compute?

No information can escape from a classical black hole: what goes in does not come out. But the quantum mechanical picture of a black hole is different. First of all, black holes are not quite black; they radiate at the Hawking temperature T given above. In addition, the well-known statement that 'a black hole has no hair' — that is, from a distance all black holes with the same charge and angular momentum look essentially alike — is now known to be not always true^{89–91}. Finally, research in string theory^{92–94} indicates that black holes may not actually destroy the information about how they were formed, but instead process it and emit the processed information as part of the Hawking radiation as they evaporate: what goes in does come out, but in an altered form.

If this picture is correct, then black holes could in principle be 'programmed': one forms a black hole whose initial conditions encode the information to be processed, lets that information be processed by the planckian dynamics at the hole's horizon, and extracts the answer to the computation by examining the correlations in the Hawking radiation emitted when the hole evaporates. Despite our lack of knowledge of the precise details of what happens when a black hole forms and evaporates (a full account must await a more exact treatment using whatever theory of quantum gravity and matter turns out to be the correct one), we can still provide a rough estimate of how much information is processed during this computation^{95–96}. Using Page's results on the rate of evaporation of a black hole⁹⁵, we obtain a lifetime for the hole $t_{\text{life}} = G^2 m^3 / 3Chc^4$, where C is a constant that depends on the number of species of particles with a mass less than $k_B T$, where T is the temperature of the hole. For $O(10^1 - 10^2)$ such species, C is on the order of $10^{-3} - 10^{-2}$, leading to a lifetime for a black hole of mass 1 kg of $\sim 10^{-19}$ s, during which time the hole can perform $\sim 10^{32}$ operations on its $\sim 10^{16}$ bits. As the actual number of effectively massless particles at the Hawking temperature of a 1-kg black hole is likely to be considerably larger than 10^2 , this number should be regarded as an upper bound on the actual number of operations that could be performed by the hole. Although this hypothetical computation is performed at ultra-high densities and speeds, the total number of bits available to be processed is not far from the number available to current computers operating in more familiar surroundings.

benefits are to be obtained by varying the volume to which the computer is confined. Generally speaking, if the computation to be performed is highly parallelizable or requires many bits of memory, the volume of the computer should be greater and the energy available to perform the computation should be spread out evenly among the different parts of the computer. Conversely, if the computation to be performed is highly serial and requires fewer bits of memory, the energy should be concentrated in particular parts of the computer.

A good measure of the degree of parallelization in a computer is the ratio between the time it takes to communicate from one side of the computer to the other, and the average time it takes to perform a logical operation. The amount of time it takes to send a message from one side of a computer of radius R to the other is $t_{\text{com}} = 2R/c$. The average time it takes a bit to flip in the ultimate laptop is the inverse of the number of operations per bit per second calculated above: $t_{\text{flip}} = \pi \hbar S / k_B 2 \ln(2) E$. The measure of the degree of parallelization in the ultimate laptop is then

$$t_{\text{com}}/t_{\text{flip}} = k_B 4 \ln(2) R E / \pi \hbar c S \propto k_B R T / \hbar c = 2 \pi R / \lambda_T$$

That is, the amount of time it takes to communicate from one side of the computer to the other, divided by the amount of time it takes to flip a bit, is approximately equal to the ratio between the size of the

system and its thermal wavelength. For the ultimate laptop, with $2R = 10^{-1}$ m, $2E/\pi\hbar \approx 10^{51}$ operations per second, and $S/k_B \ln 2 \approx 10^{31}$ bits, $t_{\text{com}}/t_{\text{flip}} \approx 10^{10}$. The ultimate laptop is highly parallel. A greater degree of serial computation can be obtained at the cost of decreasing memory space by compressing the size of the computer or making the distribution of energy more uneven. As ordinary matter obeys the Bekenstein bound^{20–22}, $k_B RE/\hbar c S > 1/2\pi$, as the computer is compressed, $t_{\text{com}}/t_{\text{flip}} \approx k_B RE/\hbar c S$ will remain greater than one, that is, the operation will still be somewhat parallel. Only at the ultimate limit of compression — a black hole — is the computation entirely serial.

Compressing the computer allows more serial computation

Suppose that we want to perform a highly serial computation on a few bits. Then it is advantageous to compress the size of the computer so that it takes less time to send signals from one side of the computer to the other at the speed of light. As the computer gets smaller, keeping the energy fixed, the energy density inside the computer increases. As this happens, different regimes in high-energy physics are necessarily explored in the course of the computation. First the weak unification scale is reached, then the grand unification scale. Finally, as the linear size of the computer approaches its Schwarzschild radius, the Planck scale is reached (Fig. 2). (No known technology could possibly achieve such compression.) At the Planck scale, gravitational effects and quantum effects are both important: the Compton wavelength of a particle of mass m , $\lambda_C = 2\pi\hbar/mc$, is on the order of its Schwarzschild radius, $2Gm/c^2$. In other words, to describe behaviour at length scales of the size $\ell_p = \sqrt{\hbar G/c^3} = 1.616 \times 10^{-35}$ m, timescales $t_p = \sqrt{\hbar G/c^5} = 5.391 \times 10^{-44}$ s, and mass scales of $m_p = \sqrt{\hbar c/G} = 2.177 \times 10^{-8}$ kg, a unified theory of quantum gravity is required. We do not currently possess such a theory. Nonetheless, although we do not know the exact number of bits that can be registered by a 1-kg computer confined to a volume of 1 l, we do know the exact number of bits that can be registered by a 1-kg computer that has been compressed to the size of a black hole⁸⁷. This is because the entropy of a black hole has a well-defined value.

In the following discussion, I use the properties of black holes to place limits on the speed, memory space and degree of serial computation that could be approached by compressing a computer to the smallest possible size. Whether or not these limits could be attained, even in principle, is a question whose answer will have to await a unified theory of quantum gravity (see Box 2).

The Schwarzschild radius of a 1-kg computer is $R_s = 2Gm/c^2 = 1.485 \times 10^{-27}$ m. The entropy of a black hole is Boltzmann's constant multiplied by its area divided by 4, as measured in Planck units. Accordingly, the amount of information that can be stored in a black hole is $I = 4\pi Gm^2/\ln(2)\hbar c = 4\pi m^2/\ln(2)m_p^2$. The amount of information that can be stored by the 1-kg computer in the black-hole limit is 3.827×10^{16} bits. A computer compressed to the size of a black hole can perform 5.4258×10^{50} operations per second, the same as the 1-l computer.

In a computer that has been compressed to its Schwarzschild radius, the energy per bit is $E/I = mc^2/I = \ln(2)\hbar c^3/4\pi mG = \ln(2)k_B T/2$, where $T = (\partial S/\partial E)^{-1} = \hbar c/4\pi k_B R_s$ is the temperature of the Hawking radiation emitted by the hole. As a result, the time it takes to flip a bit on average is $t_{\text{flip}} = \pi\hbar I/2E = \pi^2 R_s/c \ln 2$. In other words, according to a distant observer, the amount of time it takes to flip a bit, t_{flip} , is on the same order as the amount of time $t_{\text{com}} = \pi R_s/c$ it takes to communicate from one side of the hole to the other by going around the horizon: $t_{\text{com}}/t_{\text{flip}} = \ln 2/\pi$. In contrast to computation at lesser densities, which is highly parallel, computation at the horizon of a black hole is highly serial: every bit is essentially connected to every other bit over the course of a single logic operation. As noted above, the serial nature of computation at the black-hole limit can be deduced from the fact that black holes attain the Bekenstein bound^{20–22}, $k_B RE/\hbar c S = 1/2\pi$.

Constructing ultimate computers

Throughout this entire discussion of the physical limits to computation, no mention has been made of how to construct a computer that



Figure 2 Computing at the black-hole limit. The rate at which the components of a computer can communicate is limited by the speed of light. In the ultimate laptop, each bit can flip $\sim 10^{19}$ times per second, whereas the time taken to communicate from one side of the 1-l computer to the other is on the order of 10^9 s — the ultimate laptop is highly parallel. The computation can be speeded up and made more serial by compressing the computer. But no computer can be compressed to smaller than its Schwarzschild radius without becoming a black hole. A 1-kg computer that has been compressed to the black-hole limit of $R_s = 2Gm/c^2 = 1.485 \times 10^{-27}$ m can perform 5.4258×10^{50} operations per second on its $I = 4\pi Gm^2/\ln(2)\hbar c = 3.827 \times 10^{16}$ bits. At the black-hole limit, computation is fully serial: the time it takes to flip a bit and the time it takes a signal to communicate around the horizon of the hole are the same.

operates at those limits. In fact, contemporary quantum ‘microcomputers’ such as those constructed using NMR^{76–80} do indeed operate at the limits of speed and memory space described above. Information is stored on nuclear spins, with one spin registering one bit. The time it takes a bit to flip from a state $|\uparrow\rangle$ to an orthogonal state $|\downarrow\rangle$ is given by $\pi\hbar/2\mu B = \pi\hbar/2E$, where μ is the spin’s magnetic moment, B is the magnetic field, and $E = \mu B$ is the average energy of interaction between the spin and the magnetic field. To perform a quantum logic operation between two spins takes a time $\pi\hbar/2E_{\gamma}$, where E_{γ} is the energy of interaction between the two spins.

Although NMR quantum computers already operate at the limits to computation set by physics, they are nonetheless much slower and process much less information than the ultimate laptop described above. This is because their energy is locked up largely in mass, thereby limiting both their speed and their memory. Unlocking this energy is of course possible, as a thermonuclear explosion indicates. But controlling such an 'unlocked' system is another question. In discussing the computational power of physical systems in which all energy is put to use, I assumed that such control is possible in principle, although it is certainly not possible in current practice. All current designs for quantum computers operate at low energy levels and temperatures, exactly so that precise control can be exerted on their parts.

As the above discussion of error correction indicates, the rate at which errors can be detected and rejected to the environment by error-correction routines places a fundamental limit on the rate at which errors can be committed. Suppose that each logical operation performed by the ultimate computer has a probability ϵ of being erroneous. The total number of errors committed by the ultimate computer per second is then $2\epsilon E/\pi\hbar$. The maximum rate at which information can be rejected to the environment is, up to a geometric factor, $\ln(2)cS/R$ (all bits in the computer moving outward at the speed of light). Accordingly, the maximum error rate that the ultimate computer can tolerate is $\epsilon \leq \pi \ln(2)\hbar c S/2ER = 2t_{\text{flip}}/t_{\text{com}}$. That is, the maximum error rate that can be tolerated by the ultimate computer is the inverse of its degree of parallelization.

Suppose that control of highly energetic systems were to become possible. Then how might these systems be made to compute? As an example of a 'computation' that might be performed at extreme conditions, consider a heavy-ion collision that takes place in the heavy-ion collider at Brookhaven (S. H. Kahana, personal communication). If one collides 100 nucleons on 100 nucleons (that is, two nuclei with 100 nucleons each) at 200 GeV per nucleon, the operation time is $\pi\hbar/2E \approx 10^{-29}$ s. The maximum entropy can be estimated to be $\sim 4k_B$ per relativistic pion (to within a factor of less than 2 associated with the overall entropy production rate per meson), and there are $\sim 10^4$ relativistic pions per collision. Accordingly, the total amount of memory space available is $S/k_B \ln 2 \approx 10^4$ – 10^5 bits. The collision time is short: in the centre-of-mass frame the two nuclei are Lorentz-contracted to D/γ where $D = 12$ – 13 fermi and $\gamma = 100$, giving a total collision time of $\sim 10^{-25}$ s. During the collision, then, there is time to perform approximately 10^4 operations on 10^4 bits — a relatively simple computation. (The fact that only one operation per bit is performed indicates that there is insufficient time to reach thermal equilibrium, an observation that is confirmed by detailed simulations.) The heavy-ion system could be programmed by manipulating and preparing the initial momenta and internal nuclear states of the ions. Of course, we would not expect to be able to do word processing on such a 'computer'. Rather it would be used to uncover basic knowledge about nuclear collisions and quark–gluon plasmas: in the words of Heinz Pagels, the plasma 'computes itself'⁸⁸.

At the greater extremes of a black-hole computer, I assumed that whatever theory (for example, string theory) turns out to be the correct theory of quantum matter and gravity, it is possible to prepare initial states of such systems that causes their natural time evolution to carry out a computation. What assurance do we have that such preparations exist, even in principle?

Physical systems that can be programmed to perform arbitrary digital computations are called computationally universal. Although computational universality might at first seem to be a stringent demand on a physical system, a wide variety of physical systems — ranging from nearest-neighbour Ising models⁵² to quantum electrodynamics⁸⁴ and conformal field theories (M. Freedman, unpublished results) — are known to be computationally universal^{51–53,55–65}. Indeed, computational universality seems to be the rule rather than the exception. Essentially any quantum system that admits controllable nonlinear interactions can be shown to be computationally

universal^{60,61}. For example, the ordinary electrostatic interaction between two charged particles can be used to perform universal quantum logic operations between two quantum bits. A bit is registered by the presence or absence of a particle in a mode. The strength of the interaction between the particles, e^2/r , determines the amount of time $t_{\text{flip}} = \pi\hbar r/2e^2$ it takes to perform a quantum logic operation such as a controlled-NOT on the two particles. The time it takes to perform such an operation divided by the amount of time it takes to send a signal at the speed of light between the bits $t_{\text{com}} = r/c$ is a universal constant, $t_{\text{flip}}/t_{\text{com}} = \pi\hbar c/2e^2 = \pi/2\alpha$, where $\alpha = e^2/\hbar c \approx 1/137$ is the fine structure constant. This example shows the degree to which the laws of physics and the limits to computation are entwined.

In addition to the theoretical evidence that most systems are computationally universal, the computer on which I am writing this article provides strong experimental evidence that whatever the correct underlying theory of physics is, it supports universal computation. Whether or not it is possible to make computation take place in the extreme regimes envisaged in this paper is an open question. The answer to this question lies in future technological development, which is difficult to predict. If, as seems highly unlikely, it is possible to extrapolate the exponential progress of Moore's law into the future, then it will take only 250 years to make up the 40 orders of magnitude in performance between current computers that perform 10^{10} operations per second on 10^{10} bits and our 1-kg ultimate laptop that performs 10^{51} operations per second on 10^{31} bits. □

1. Maxwell, J. C. *Theory of Heat* (Appleton, London, 1871).
2. Smoluchowski, F. *Vorträge über die kinetische Theorie der Materie u. Elektrizität* (Leipzig, 1914).
3. Szilard, L. Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen. *Z. Physik* **53**, 840–856 (1929).
4. Brillouin, L. *Science and Information Theory* (Academic Press, New York, 1953).
5. Landauer, R. Irreversibility and heat generation in the computing process. *IBM J. Res. Dev.* **5**, 183–191 (1961).
6. Keyes, R. W. & Landauer, R. Minimal energy dissipation in logic. *IBM J. Res. Dev.* **14**, 152–157 (1970).
7. Landauer, R. Dissipation and noise-immunity in computation and communication. *Nature* **335**, 779–784 (1988).
8. Landauer, R. Information is physical. *Phys. Today* **44**, 23–29 (1991).
9. Landauer, R. The physical nature of information. *Phys. Lett. A* **217**, 188–193 (1996).
10. von Neumann, J. *Theory of Self-Reproducing Automata* Lect. 3 (Univ. Illinois Press, Urbana, IL, 1966).
11. Lebedev, D. S. & Levitin, L. B. Information transmission by electromagnetic field. *Inform. Control* **9**, 1–22 (1966).
12. Levitin, L. B. in *Proceedings of the 3rd International Symposium on Radio Electronics* part 3, 1–15 (Varna, Bulgaria, 1970).
13. Levitin, L. B. Physical limitations of rate, depth, and minimum energy in information processing. *Int. J. Theor. Phys.* **21**, 299–309 (1982).
14. Levitin, L. B. Energy cost of information transmission (along the path to understanding). *Physica D* **120**, 162–167 (1998).
15. Margolus, N. & Levitin, L. B. in *Proceedings of the Fourth Workshop on Physics and Computation—PhysComp96* (eds Toffoli, T., Biale, M. & Leão, J.) (New England Complex Systems Institute, Boston, MA, 1996).
16. Margolus, N. & Levitin, L. B. The maximum speed of dynamical evolution. *Physica D* **120**, 188–195 (1998).
17. Bremermann, H. J. in *Self-Organizing Systems* (eds Yovits, M. C., Jacobi, G. T. & Goldstein, G. D.) 93–106 (Spartan Books, Washington DC, 1962).
18. Bremermann, H. J. in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (eds LeCam, L. M. & Neyman, J.) Vol. 4, 15–20 (Univ. California Press, Berkeley, CA, 1967).
19. Bremermann, H. J. Minimum energy requirements of information transfer and computing. *Int. J. Theor. Phys.* **21**, 203–217 (1982).
20. Bekenstein, J. D. Universal upper bound on the entropy-to-energy ration for bounded systems. *Phys. Rev. D* **23**, 287–298 (1981).
21. Bekenstein, J. D. Energy cost of information transfer. *Phys. Rev. Lett.* **46**, 623–626 (1981).
22. Bekenstein, J. D. Entropy content and information flow in systems with limited energy. *Phys. Rev. D* **30**, 1669–1679 (1984).
23. Aharonov, Y. & Bohm, D. Time in the quantum theory and the uncertainty relation for the time and energy domain. *Phys. Rev.* **122**, 1649–1658 (1961).
24. Aharonov, Y. & Bohm, D. Answer to Fock concerning the time-energy indeterminacy relation. *Phys. Rev. B* **134**, 1417–1418 (1964).
25. Anandan, J. & Aharonov, Y. Geometry of quantum evolution. *Phys. Rev. Lett.* **65**, 1697–1700 (1990).
26. Peres, A. *Quantum Theory: Concepts and Methods* (Kluwer, Hingham, 1995).
27. Lecerf, Y. Machines de Turing réversibles. *C.R. Acad. Sci.* **257**, 2597–2600 (1963).
28. Bennett, C. H. Logical reversibility of computation. *IBM J. Res. Dev.* **17**, 525–532 (1973).
29. Bennett, C. H. Thermodynamics of computation—a review. *Int. J. Theor. Phys.* **21**, 905–940 (1982).
30. Bennett, C. H. Demons, engines and the second law. *Sci. Am.* **257**, 108 (1987).
31. Fredkin, E. & Toffoli, T. Conservative logic. *Int. J. Theor. Phys.* **21**, 219–253 (1982).
32. Likharev, K. K. Classical and quantum limitations on energy consumption in computation. *Int. J. Theor. Phys.* **21**, 311–325 (1982).
33. Seitz, C. L. et al. in *Proceedings of the 1985 Chapel Hill Conference on VLSI* (ed. Fuchs, H.) (Computer Science Press, Rockville, MD, 1985).

34. Merkle, R. C. Reversible electronic logic using switches. *Nanotechnology* **34**, 21–40 (1993).
35. Younis, S. G. & Knight, T. F. in *Proceedings of the 1993 Symposium on Integrated Systems*, Seattle, Washington (eds Berriello, G. & Ebeling, C.) (MIT Press, Cambridge, MA, 1993).
36. Lloyd, S. & Pagels, H. Complexity as thermodynamic depth. *Ann. Phys.* **188**, 186–213 (1988).
37. Lloyd, S. Use of mutual information to decrease entropy—implications for the Second Law of Thermodynamics. *Phys. Rev. A* **39**, 5378–5386 (1989).
38. Zurek, W. H. Thermodynamic cost of computation, algorithmic complexity and the information metric. *Nature* **341**, 119–124 (1989).
39. Leff, H. S. & Rex, A. F. *Maxwell's Demon: Entropy, Information, Computing* (Princeton Univ. Press, Princeton, 1990).
40. Lloyd, S. Quantum mechanical Maxwell's demon. *Phys. Rev. A* **56**, 3374–3382 (1997).
41. Benioff, P. The computer as a physical system: a microscopic quantum mechanical Hamiltonian model of computers as represented by Turing machines. *J. Stat. Phys.* **22**, 563–591 (1980).
42. Benioff, P. Quantum mechanical models of Turing machines that dissipate no energy. *Phys. Rev. Lett.* **48**, 1581–1585 (1982).
43. Feynman, R. P. Simulating physics with computers. *Int. J. Theor. Phys.* **21**, 467 (1982).
44. Feynman, R. P. Quantum mechanical computers. *Optics News* **11**, 11 (1985); reprinted in *Found. Phys.* **16**, 507 (1986).
45. Zurek, W. H. Reversibility and stability of information-processing systems. *Phys. Rev. Lett.* **53**, 391–394 (1984).
46. Peres, A. Reversible logic and quantum computers. *Phys. Rev. A* **32**, 3266–3276 (1985).
47. Deutsch, D. Quantum-theory, the Church-Turing principle, and the universal quantum computer. *Proc. R. Soc. Lond. A* **400**, 97–117 (1985).
48. Margolus, N. Quantum computation. *Ann. N.Y. Acad. Sci.* **480**, 487–497 (1986).
49. Deutsch, D. Quantum computational networks. *Proc. R. Soc. Lond. A* **425**, 73–90 (1989).
50. Margolus, N. in *Complexity, Entropy, and the Physics of Information*, Santa Fe Institute Studies in the Sciences of Complexity Vol. VIII (ed. Zurek, W. H.) 273–288 (Addison Wesley, Redwood City, 1991).
51. Lloyd, S. Quantum-mechanical computers and uncomputability. *Phys. Rev. Lett.* **71**, 943–946 (1993).
52. Lloyd, S. A potentially realizable quantum computer. *Science* **261**, 1569–1571 (1993).
53. Lloyd, S. Necessary and sufficient conditions for quantum computation. *J. Mod. Opt.* **41**, 2503–2520 (1994).
54. Shor, P. in *Proceedings of the 35th Annual Symposium on Foundations of Computer Science* (ed. Goldwasser, S.) 124–134 (IEEE Computer Society, Los Alamitos, CA, 1994).
55. Lloyd, S. Quantum-mechanical computers. *Sci. Am.* **273**, 140–145 (1995).
56. DiVincenzo, D. Quantum computation. *Science* **270**, 255–261 (1995).
57. DiVincenzo, D. P. 2-Bit gates are universal for quantum computation. *Phys. Rev. A* **51**, 1015–1022 (1995).
58. Sleator, T. & Weinfurter, H. Realizable universal quantum logic gates. *Phys. Rev. Lett.* **74**, 4087–4090 (1995).
59. Barenco, A. *et al.* Elementary gates for quantum computation. *Phys. Rev. A* **52**, 3457–3467 (1995).
60. Lloyd, S. Almost any quantum logic gate is universal. *Phys. Rev. Lett.* **75**, 346–349 (1995).
61. Deutsch, D., Barenco, A. & Ekert, A. Universality in quantum computation. *Proc. R. Soc. Lond. A* **449**, 669–677 (1995).
62. Cirac, J. I. & Zoller, P. Quantum computation with cold ion traps. *Phys. Rev. Lett.* **74**, 4091–4094 (1995).
63. Pellizzari, T., Gardiner, S. A., Cirac, J. I. & Zoller, P. Decoherence, continuous observation, and quantum computing—a cavity QED model. *Phys. Rev. Lett.* **75**, 3788–3791 (1995).
64. Turchette, Q. A., Hood, C. J., Lange, W., Mabuchi, H. & Kimble, H. J. Measurement of conditional phase-shifts for quantum logic. *Phys. Rev. Lett.* **75**, 4710–4713 (1995).
65. Monroe, C., Meekhof, D. M., King, B. E., Itano, W. M. & Wineland, D. J. Demonstration of a fundamental quantum logic gate. *Phys. Rev. Lett.* **75**, 4714–4717 (1995).
66. Grover, L. K. in *Proceedings of the 28th Annual ACM Symposium on the Theory of Computing* 212–218 (ACM Press, New York, 1996).
67. Lloyd, S. Universal quantum simulators. *Science* **273**, 1073–1078 (1996).
68. Zalka, C. Simulating quantum systems on a quantum computer. *Proc. R. Soc. Lond. A* **454**, 313–322 (1998).
69. Shor, P. W. A scheme for reducing decoherence in quantum memory. *Phys. Rev. A* **52**, R2493–R2496 (1995).
70. Steane, A. M. Error correcting codes in quantum theory. *Phys. Rev. Lett.* **77**, 793–797 (1996).
71. Laflamme, R., Miquel, C., Paz, J. P. & Zurek, W. H. Perfect quantum error correcting code. *Phys. Rev. Lett.* **77**, 198–201 (1996).
72. DiVincenzo, D. P. & Shor, P. W. Fault-tolerant error correction with efficient quantum codes. *Phys. Rev. Lett.* **77**, 3260–3263 (1996).
73. Shor, P. in *Proceedings of the 37th Annual Symposium on the Foundations of Computer Science* 56–65 (IEEE Computer Society Press, Los Alamitos, CA, 1996).
74. Preskill, J. Reliable quantum computers. *Proc. R. Soc. Lond. A* **454**, 385–410 (1998).
75. Knill, E., Laflamme, R. & Zurek, W. H. Resilient quantum computation. *Science* **279**, 342–345 (1998).
76. Cory, D. G., Fahmy, A. F. & Havel, T. F. in *Proceedings of the Fourth Workshop on Physics and Computation—PhysComp96* (eds Toffoli, T., Biafore, M. & Leão, J.) 87–91 (New England Complex Systems Institute, Boston, MA, 1996).
77. Gershenfeld, N. A. & Chuang, I. L. Bulk spin-resonance quantum computation. *Science* **275**, 350–356 (1997).
78. Chuang, I. L., Vandersypen, L. M. K., Zhou, X., Leung, D. W. & Lloyd, S. Experimental realization of a quantum algorithm. *Nature* **393**, 143–146 (1998).
79. Jones, J. A., Mosca, M. & Hansen, R. H. Implementation of a quantum search algorithm on a quantum computer. *Nature* **393**, 344–346 (1998).
80. Chuang, I. L., Gershenfeld, N. & Kubinec, M. Experimental implementation of fast quantum searching. *Phys. Rev. Lett.* **80**, 3408–3411 (1998).
81. Kane, B. A silicon-based nuclear-spin quantum computer. *Nature* **393**, 133 (1998).
82. Nakamura, Y., Pashkin, Yu. A. & Tsai, J. S. Coherent control of macroscopic quantum states in a single-Cooper-pair box. *Nature* **398**, 786–788 (1999).
83. Mooij, J. E. *et al.* Josephson persistent-current qubit. *Science* **285**, 1036–1039 (1999).
84. Lloyd, S. & Braunstein, S. Quantum computation over continuous variables. *Phys. Rev. Lett.* **82**, 1784–1787 (1999).
85. Abrams, D. & Lloyd, S. Nonlinear quantum mechanics implies polynomial-time solution for NP-complete and P problems. *Phys. Rev. Lett.* **81**, 3992–3995 (1998).
86. Zel'dovich, Ya. B. & Novikov, I. D. *Relativistic Astrophysics* (Univ. of Chicago Press, Chicago, 1971).
87. Novikov, I. D. & Frolov, V. P. *Black Holes* (Springer, Berlin, 1986).
88. Pagels, H. *The Cosmic Code: Quantum Physics as the Language of Nature* (Simon and Schuster, New York, 1982).
89. Coleman, S., Preskill, J. & Wilczek, F. Growing hair on black-holes. *Phys. Rev. Lett.* **67**, 1975–1978 (1991).
90. Preskill, J. Quantum hair. *Phys. Scr. T* **36**, 258–264 (1991).
91. Fiola, T. M., Preskill, J. & Strominger, A. Black-hole thermodynamics and information loss in 2 dimensions. *Phys. Rev. D* **50**, 3987–4014 (1994).
92. Susskind, L. & Uglum, J. Black-hole entropy in canonical quantum-gravity and superstring theory. *Phys. Rev. D* **50**, 2700–2711 (1994).
93. Strominger, A. & Vafa, C. Microscopic origin of the Bekenstein-Hawking entropy. *Phys. Lett. B* **37**, 99–104 (1996).
94. Das, S. R. & Mathur, S. D. Comparing decay rates for black holes and D-branes. *Nucl. Phys. B* **478**, 561–576 (1996).
95. Page, D. N. Particle emission rates from a black-hole: massless particles form an uncharged non-rotating black-hole. *Phys. Rev. D* **13**, 198 (1976).
96. Thorne, K. S., Zurek, W. H. & Price, R. H. in *Black Holes: The Membrane Paradigm* Ch. VIII (eds Thorne, K. S., Price, R. H. & Macdonald, D. A.) 280–340 (Yale Univ. Press, New Haven, CT, 1986).

Biotechnology selection

New ways of handling slurry, rats in transit, and DNA.

Disposable carriers

Matrix www.matrixtechcorp.com

Disposable yet reusable tube transport

Disposable carriers are now available for the complete line of Matrix Technologies storage tubes. The carriers bind together 1.4 ml Screen-Mates or 2-D bar-coded TrakMates for easy placement in a Snap or Latch Rack. The carriers can be used to refill a rack with up to 96 tubes, or to transfer a set of tubes from one rack to another. They are reusable and can also be used to enhance sample security by binding each tube to all others within a rack.

Reader Service No. 100



On the rack: a disposable approach from Matrix.



Height of comfort: Taconic's travel cages.

FuGENE

Roche Diagnostics <http://biochem.roche.com>

Been to the web site, got the T-shirt?

The FuGENE transfection reagent is now available as a multi-pack of 5 × 1 ml in addition to the 0.4 ml and standard 1 ml pack sizes. FuGENE is an optimized blend of lipids and other proprietary compounds for transfection of eukaryotic cells. Neat DNA samples can be incubated directly with diluted FuGENE reagent then added immediately to cells. Roche Diagnostics is currently running a questionnaire on its web site collecting data on the capacity of FuGENE 6 to transfect different cell lines. Researchers submitting results with a cell line that is not already listed on the web site are offered the reward of a FuGENE 6 Transfection Reagent T-shirt.

Reader Service No. 101

sterile or non-sterile, and are suitable for enzyme or receptor assays. The plate footprint is designed for automated liquid handling workstations including Zymark, Tecan, Hamilton, Rosys and Beckman.

Reader Service No. 103

Rat Transit Cage

Taconic www.taconic.com

Comfortable travel for rodents

Taconic has introduced a new high lid for the

Taconic Transit Cage (TTC) enabling it to accommodate Taconic's Murine Pathogen Free (MPF) rats. The high lid for rats complements the currently available low lid TTC for shipments of mice. The TTC was designed to provide an optimal shipping environment for lab rats and mice. Features of the new high lid TTC are polypropylene plastic construction and tear-resistant filters that prevent airborne contaminants from entering the animal chamber. Compartment dividers are available to divide the TTC into two, three or four compartments.

Spex FluoroMax-3

Jobin Yvon www.jyhoriba.co.uk

Update to an established system

The latest incarnation of the Spex FluoroMax spectrofluorimeter includes an extended range of accessories. New features include auto-calibration on start-up, computer-controlled bandpass and the capability to control the FluoroMax-3 via any PC running Windows 95/98/NT. The FluoroMax-3 fits these features into a smaller package than previous models.

Reader Service No. 102

96 Deep-Well Plate

Bel-Art www.bel-art.com

Sample size the key

The Scienceware 96 Deep-Well Plate from Bel-Art Products contains a larger sample, making it possible to do more testing from the same well. These plates contain 96 deep wells, each containing 2 ml. They are made of polypropylene and are available individually packaged, either

Custom Peptides & Polyclonal Antibodies

Please contact us about complete services for your custom peptide and polyclonal antibody needs.

Research Genetics
U.S. or Canada 800-533-4363
FAX 256-536-9016

Check out our homepage at
<http://www.resgen.com>

READER ENQUIRY NO. 39

TREVIGEN **new** **PARP**
Poly (ADP-ribose) Polymerase Reagents

- PARP inhibition assay
- Biotinylated NAD
- Recombinant human PARP
- Anti PARP and PAR

www.trevigen.com
US and Canada: 1 800 873 8443
TEL: 301 216 2800 • FAX: 301 216 2801

READER ENQUIRY NO. 43

BLOCKERS

K⁺ Channel Blocker **E4031** **AA 861** **Arachidonic Acid Cascade Blocker**

Wako USA 877-714-1921 **Wako** **Wako GmbH** (49)02131 3110
www.wako-chem.co.jp

READER ENQUIRY NO. 35

BACs • YACs • PACs
cDNAs

Research Genetics provides clones and colony membranes for a number of human, mouse and rat libraries as well as a custom screening service for the libraries.

Research Genetics
U.S. or Canada 800-533-4363
U.K. 0-800-89-1393
FAX 256-536-9016

Check out our homepage at
<http://www.resgen.com>

READER ENQUIRY NO. 40

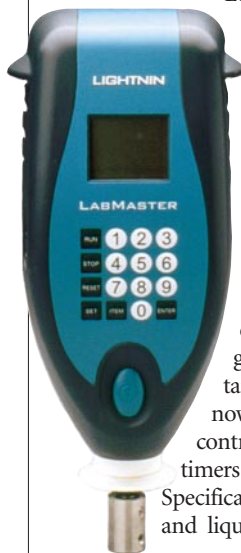
new on the market

The TTC is autoclavable, reusable and may be stacked (or nested when empty) to save space.
Reader Service No. 104

Lightnin

Nalge Nunc www.nalgenunc.com
Hands-off mixing

These new 250-r.p.m. overhead drive Biotech Mixers are designed to mate with Nalgene's Lower Assembly Mixing Systems to provide a complete sanitary mixing system to function on Nalge Nunc International Bioprocess Vessels, Closed Dome Tanks and Sanitary Conical Tanks. The tank and mixer packages with Biotech Overhead Mixers are designed as an economical alternative to glass and stainless steel tank systems. The mixers now feature a variable speed control and programmable timers for hands-off operation. Specifically designed for liquids and liquid slurries, they can be



Nalgene's Biotech Mixer.

programmed to function in either clockwise or counter-clockwise direction, and are available in 100-volt and 220-volt units.
Reader Service No. 105

PCR Master Mixes

ABgene www.abgene.com
A new format for PCR mixes

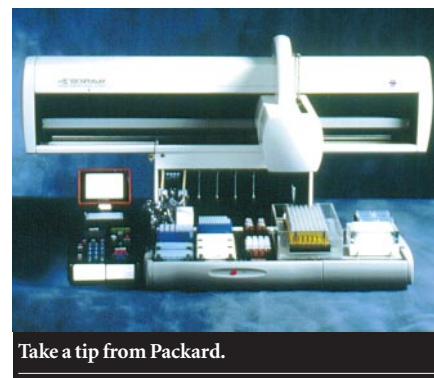
ABgene's range of PCR Master Mixes ("just add template and primers") is now available in ReddyMix format. Formulated with a red dye and precipitant for direct loading of PCR products onto agarose gels, ReddyMix Master Mix is claimed to significantly reduce both PCR set-up time, and post-PCR analysis time. No loading dye is required at the end of thermal cycling, just remove a sample and load directly on the gel. The red indicator dye has no negative effect on PCR performance. In addition to the time saving when using ReddyMix, the reduction of pipetting steps increases reproducibility and consistency of results.

Reader Service No. 106

MultiPROBE II

Packard www.packardinstrument.com
Automated DNA processing

Further integration modules, designed to improve performance and flexibility in genomic applications, have been introduced for the MultiPROBE II robotic liquid han-



Take a tip from Packard.

dling system. The system can be configured to automate all liquid handling steps involved in processing DNA extraction and plasmid preparation kits using vacuum control methods. The system can be configured to automate all liquid handling steps in cycle sequencing reactions. The new options for the MultiPROBE II are a Low Volume Tip to allow dispensing using standard syringe technology to volumes as low as 200 nL, though it can also be used to dispense volumes as high as 250 µL or more. And a Six Way External Valve option allows several different system fluids to be delivered.

Reader Service No. 107

These notes are compiled in the Nature office from information provided by the manufacturers.

ADVERTISEMENTS

www.alphainnotech.com

Alpha Innotech Corporation

READER ENQUIRY NO. 48

Cross species Primer design with Primer Premier

- ✓ Allele specific primers
- ✓ Degenerate Primers
- ✓ ClustalW alignment
- ✓ Pathogenic primers
- ✓ Restriction enzyme & motif analysis
- ✓ Nested/multiplex primers and more...

Download free demo from
<http://www.PremierBiosoft.com>

Call 888-847-7494 (Toll free in US)
650-856-2703 (International)

PREMIER Biosoft International
Software solutions for the molecular biologist

READER ENQUIRY NO. 45

Look no further

for the best value in

Oligo Synthesis

Competitive pricing
Routine 24 hour dispatch • Every Oligo QC verified by PAGE
Rigorous batch QC by PCR and Mass Spec
Web ordering and tracking • Local Freephone numbers
A great deal in any language!!

SIGMA GENOSYS (+44) (0) 1223 839000 • fax (+44) (0) 1223 839200
email: info@sigma-genosys.co.uk • www.sigma-genosys.co.uk

READER ENQUIRY NO. 33

Diagnostic Reagents

Pefachrome®-Series: chromogenic and fluorigenic substrates

Pefabloc®-Series: serine proteinase inhibitors

Purified snake venom proteins (e.g. Protac®, Ecarin, Batroxobin)

PENTAPHARM LTD
CH 4002 Basel Fax: +41 61 319 96 19
<http://www.pentapharm.ch>

READER ENQUIRY NO. 16

CometAssay™
Single Cell Gel Electrophoresis

www.trevigen.com **1-800-873-8443**

READER ENQUIRY NO. 47

REAGENT KITS

NEW FLUOROSPARK PTP KIT
Simple, Rapid and Sensitive
Quantifies PTP activity

ESTROGEN RECEPTOR KIT
High throughput screening of endocrine disrupting chemicals

SLP REAGENT KIT
Highly sensitive peptidoglycan and (1-3)-β-glucan Assay

Wako Wako USA 877-714-1921 Wako GmbH (49)02131 3110
www.wako-chem.co.jp

READER ENQUIRY NO. 42